Title
What Makes Reading Difficult? An Investigation of the Contribution of Passage, Task, and Reader Characteristics on Item Difficulty, Using Explanatory Item Response Models

Author
Toyama, Yukie

Publication Date
2019

Peer reviewed|Thesis/dissertation

What Makes Reading Difficult? An Investigation of the Contribution of Passage, Task, and Reader Characteristics on Item Difficulty, Using Explanatory Item Response Models

By

Yukie Toyama

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor P. David Pearson
Professor Susanne Gahl
Dr. Elfrieda H. Hiebert

Summer 2019

What Makes Reading Difficult? An Investigation of the Contribution of Passage, Task, and Reader Characteristics on Item Difficulty, Using Explanatory Item Response Models

Abstract

What Makes Reading Difficult? An Investigation of the Contribution of Passage, Task, and Reader Characteristics on Item Difficulty, Using Explanatory Item Response Models

by

Yukie Toyama

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

Reading comprehension (RC) is often viewed as a multi-faceted, multi-layered construct (Broek & Espin, 2012; Duke, 2005; Graesser & McNamara, 2011; Perfetti & Stafura, 2014), which manifests through complex interactions among three broad factors: the reader, the passage, and the task, all residing in a particular socio-cultural context (RAND Reading Study Group, 2002). Drawing on the item difficulty modeling paradigm, this study examined how these three factors as well as their interactions affected comprehension difficulty. Specifically, the study used explanatory item response models (De Boeck & Wilson, 2004) to analyze a vertically-scaled item response matrix from an operational online assessment, which included a wide range of readers (n=10,547) as well as of passages (n=48), covering grades 1 through 12+. Analyses indicated that it is text features, as measured by computational text analyzers, rather than task features as coded by human raters, that explained over half the variance in item difficulty, after controlling for student general vocabulary knowledge. Specifically, sentence length, word frequency, syntactic simplicity, and temporality (i.e., the extent to which the text has time markers) were found to significantly affect comprehension difficulty in both model building and cross validation analyses. Further, small but significant interaction effects were found, indicating that these textual effects were moderated by student general vocabulary knowledge as well as task demands as captured by item types. In general, readers with higher vocabulary knowledge benefitted more from traditional textual affordances (e.g., shorter sentences, familiar words) than their peers with lower vocabulary knowledge, especially when questions asked them to recall specific localized information without accessing the source passage. However, a reverse effect was found with temporality: passages with more time markers helped low vocabulary readers, while it was low temporality passages that helped high vocabulary readers. The implications of these findings as well as their limitations are discussed as they relate to the measurement of RC and to instructional practice.

To those who guided and supported me along the way.

Table of Contents

List of Tables

List of Figures

Acknowledgement

I am deeply grateful to many people who guided and supported me to complete this dissertation. Although only a small subset of people are mentioned here, I am truly grateful to everyone who helped me along the way.

I thank Mark Wilson for his technical advice and critical feedback throughout my Ph.D training, always pushing me make the best efforts in contributing to the fields of measurement and science. I thank P. David Pearson for his mentorship and generosity with his time and resource, which greatly helped me appreciate the complexities involved in the act of reading. I thank Freddy Hiebert for her encouragement to pursue a research agenda on text complexity and for opportunities to network with like-minded researchers in her circle. I thank Susanne Gahl for sharing her expertise in psycholinguistics and quantitative modeling.

I would also like to acknowledge Alex Spichtig and her team at ReadingPlus for giving me access to their assessment data. I also would like to thank Jeff Elmore at MetaMetrics, Arthur Graesser at University of Memphis, and Kathleen Sheehan at ETS, for running a batch analysis of assessment passages used in my dissertation with their respective text analysis tools.

I thank the members of the Quantitative Methods and Evaluation group at the UC Berkeley Graduate School of Education. I specially thank JinHo Kim, Diah Wihardini, Dan Furr, and Perman Gochyyev, for providing technical and emotional support when I needed them the most. I also thank the members of P.David Pearson's research group, especially Catherine Miller, for providing insights and feedback to the earlier versions of my work.

Lastly but by no means least, I thank my husband and daughter, Eraj and Saiyra Siddiqui, for their eternal patience and cheering. I would never have completed my dissertation without their support.

Chapter 1. Introduction

Most reading researchers agree that reading comprehension (RC) is a multi-faceted, multi-layered construct (Broek & Espin, 2012; Duke, 2005; Graesser & McNamara, 2011; Perfetti & Stafura, 2014), which manifests through the complex interaction among three broad factors: the reader, the passage, and the task, all residing in a particular socio-cultural context (RAND Reading Study Group, 2002). Investigating this complexity deepens our understanding about contributions of the reader, the passage, and the task, as well as their complex interactions in the face of a challenge to comprehension.

While the three-factor view of reading comprehension appears to be widely endorsed by reading researchers, surprisingly little attempt has been made to directly model the phenomena represented in this view. Rather, the great majority of quantitative research in reading research has focused on cognitive and affective factors and their interrelations within the reader that underlie reading comprehension. For example, one line of such research has demonstrated a strong correlation between foundational early literacy skills such as oral reading fluency and later reading comprehension outcome (e.g., Kendeou, van den Broek, White, & Lynch, 2009; Oakhill & Cain, 2012). In this research, modeling is typically done by aggregating raw scores at the test level to construct variables, and their predictive relationships are investigated through multiple regression, structural equation modeling, or path analysis. As such, this research does not explicitly model the empirical phenomena that a student reads a particular text to answer a particular question. Nor does it model how particular features of a passage or item contribute to student's reading comprehension performance.

Another line of quantitative reading research has focused on the readability of text in an effort to develop regression formulas to estimate passage difficulty, in order, ultimately, to match a "just right" text to individual readers (e.g., Bormuth, 1969; Klare, 1984). In this research, various linguistic and discourse characteristics of text are used to predict a pre-determined measure of text difficulty, typically determined by students' RC performance, experts' judgments, and/or the publishers' grade level determination. Naturally, text and its linguistic features have been the central focus of the readability research, while task features (e.g., types of RC processing, purpose of reading) and reader characteristics (e.g., readers' interest, background knowledge, memory, attention) have been largely ignored. In fact, some readability researchers sought to eliminate variance caused by non-textual factors (e.g., tasks, test writers, and scorers) by using a particular test format known as the Cloze procedure (Taylor, 1953). The Cloze procedure is a commonly used test format that deletes certain words from a passage at equal intervals (e.g., every fifth word) and asks students to fill in (the classic Cloze procedure) the word or select it from multiple-choice options (a later adaptation dubbed the Maze procedure). Proponents (e.g., Bormuth, 1968; Gellert & Elbro, 2013) argue that the Cloze method is an efficient measure of RC, eliminating the non-textual effects. However, it has been criticized as not being able to measure comprehension beyond a single sentence (e.g., Shanahan, Kamil, & Tobin, 1982). Additionally, studies have shown that various test and item features affect students' Cloze performance; these features include the type of words dropped (e.g., content vs. function words), reduction rates, the amount/type of context required for closure, and the response method (Abraham & Chapelle, 1992; Bachman, 2006; Kobayashi, 2003). Taken together, the readability research falls short of adequately modeling non-textual factors that account for comprehension difficulty.

Recognizing these gaps in the literature, this study used explanatory item response models (De Boeck & Wilson, 2004) to investigate the effects of passage, item and student characteristics as well as their interactions on students' comprehension of informational text. Specifically, the study examines item response data from a computer-adaptive RC assessment program, which included a wide range of readers and passages. An anchoring design was devised to link response data via common items and to place them onto a common vertical scale. The questions guiding this effort were: (a) which passage and task features best predict the difficulty of RC items after controlling for the reader's general vocabulary level, and (b) whether the effects of particular passage features were moderated by task and reader characteristics.

Chapter 2. Literature Review

To situate the current study in the existing literature and develop theoretical and methodological frameworks, this chapter reviews relevant literature on the following four topics:

- the RAND heuristic for reading comprehension,
- item difficulty modeling in reading comprehension research,
- quantitative analytical tools of text complexity, and
- explanatory item response models.

Materials reviewed in the following sections, in concert, will form the background information for the study.

The RAND Heuristic for Reading Comprehension

In the late 1990s, the RAND Reading Study Group (RRSG), a federally funded expert panel, was tasked to develop a research agenda to address the most-pressing issues in literacy. In response, the RRSG came up with a vision of the proficient adult reader as someone who can comprehend a variety of texts for a variety of purposes, even when the material is not easily comprehensible or inherently interesting. The panel went on to define reading comprehension as "the process of simultaneously extracting and constructing meaning through interaction and involvement with written language" RRSG, 2002, p. 11)", which involves the following three primary elements:

- the reader who does the comprehending; this element includes cognitive capabilities (e.g., attention, memory), motivation (e.g., interest, efficacy as reader), knowledge (e.g., vocabulary, background, reading strategies), and experience that the reader brings to the act of reading;
- the text that is to be comprehended; this is broadly defined to include any print or electronic text; and
- the activity in which comprehension is part; this element includes purposes, processes, and consequences of reading (e.g., knowledge gain, engagement with the text).

These elements are depicted in a popular diagram shown in Figure 1. Interestingly, the RRSG called this a "heuristic", rather than a model, whose purpose is to show the interrelatedness of the three elements.

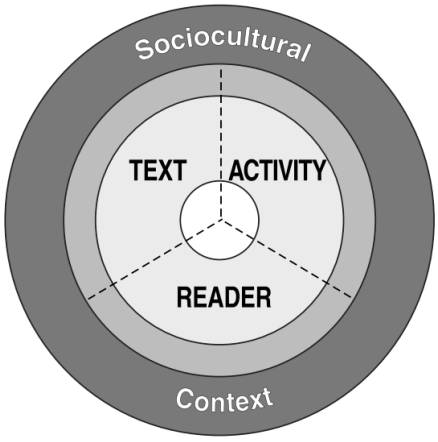

Figure 1. RAND heuristic of reading comprehension

The diagram also indicates how the three primary elements reside in a particular sociocultural context that shapes the identity and capabilities of the reader, the availability and value of the texts, and the reading activities, while at the same time the three elements influence the context. From this perspective, the context goes well beyond the classroom, including communities in which students live and learn to read.

The RRGS' definition makes it clear that the text is not the sole determinant of reading comprehension even if it plays an important role as a source from which the reader extracts meaning. The reader also has to develop various representations of the text through the reading activities that involve interacting with the written language as well as their prior knowledge.

Item Difficulty Modeling in Reading Comprehension Research

Apart from the readability research (see Klare 1984 for review of this research), there exists only a handful of large-scale empirical studies that investigated the contribution of various passage and item features to students' RC performance.[1] One of the earliest studies by Drum, Calfee, & Cook (1981), developed a framework for RC multiple-choice tests that identified four structural components of the RC assessments: (1) passage, (2) question stem, (3) correct answer, and (4) distractors— these are the column headers in Figure 2 below. The framework further specified three categories of item feature variables that were implicated in the literature to affect student RC performance: (a) word recognition/meaning, (b) knowledge of syntactic/semantic relationships, and (c) test formats—these are the row headers in the figure. Listed in each cell in the figure are the quantifiable passage and item features investigated in Drum et al.'s study.

Predictor Types	**Structural Components**			
	1. Passage	**2. Question-stem**	**3. Correct answer**	**4. Distractors**
a. Word recognition / meaning	log of number of unique content words % content words	% content words % new content words % non-Dale-Chall words	% content words % new content words % non-Dale-Chall words	% content words % new content words % non-Dale-Chall words
b. Syntactic / semantic forms	% content-function words avg. sentence length	% content-function words	NA	NA
c. Test format	NA	NA	need info external to passage	one or more distracters are plausible

Figure 2. Passage and item predictors used in Drum et al., (1981)

Guided by this framework, Drum et al. examined the effects of the four structural test elements on observed difficulty of 210 RC items from three standardized test programs designed for grades 1-12. Through separate stepwise multiple regression analyses for 18 grade-by-test form combinations, the study found that the predictor variables related to incorrect answer

[1] Small-scale experimental studies exist that examined the effect of a few particular text features (e.g., cohesion, word frequency) on reading comprehension by manipulating the text features. For the review of some of these studies, see Amendum, Conradi, & Hiebert (2018) and McNamara et al.(2010).

choices, particularly the plausibility of distractors, explained the largest variance in the observed item difficulties. The remaining three structural components, namely the passage, the question stem, and the correct answer, equally contributed to the unexplained variance left by the incorrect answer-choice predictors.

Drums et al.'s study also indicated that the passage and item features have differential effects on item difficulty depending on students' developmental levels. For example, predictors that reflect word recognition and word meaning in passages (i.e., unique words and proportion of content words in passages) tended to make RC items relatively more difficult for younger readers than for older readers. Similarly, the ratio of content-function words (e.g., *Wh*-words, pronouns, and conjunctions) in the question stem to those in the source passage— a measure of syntactic complexity of a question stem— consistently depressed performance of younger readers, while this same feature improved the performance of older students. Further, the implausibility of one or more incorrect choices tended to have a large negative effect on the item difficulty (i.e., making the items easier), especially for older students.

While Drums et al.'s findings shed some light on developmental differences in the contributions of passage and item features to student performance, they were grossly generalized patterns observed descriptively in the results from the 18 separate multiple regressions. As such, these patterns did not always hold true for all grade-by-test form combinations examined. Further, the authors did not explicitly model student grade level in their regression analyses, thus its interactions with the passage and item feature predictors were not statistically investigated. Such an analysis would require linking separate test forms for different grades within and across test programs on a common vertical scale.

While Drum et al.'s framework identified the structural test components that affect item difficulty, Embretson & Wetzel (1987) expanded the effort by proposing a model that specified cognitive processes involved in responding to multiple-choice items designed for older readers. The model draws on Kintsch's Coherence-Integration (CI) theory (Kintsch, 1988, 1994; Kintsch & van Dijk, 1978), which specifies processes associated with encoding, coherence, and integration. The top-most panel in Figure 3 outlines Embretson and Wetzel's model, which consists of two stages: (1) Text Representation and (2) Response Decision. The former involves the reader encoding visual information from the passage into a meaningful mental representation. Word frequency, an indicator of vocabulary demand of a text, was one of the text characteristics that were hypothesized to affect this process (i.e., if the text includes more frequently-used words, it is easier to process). The second step within the Text Representation stage is the Coherence Process, in which the reader supplies information from memory and background knowledge to integrate the text representation into their own cognitive network. Drawing on Kintch and Van Dijk (1978), Embretson and Wentzel hypothesized that the density of propositions (or idea units) would affect the Coherence Process (i.e., more propositionally-dense the text is, more ideas need to be stored and integrated, thus posing more demand in text processing.)

The second stage in Embretson and Wetzel's model, Response Decision, captures the reader selecting the correct answer choice after having read the question and the answer choices, and having compared the alternative choices against the source passage. Within this stage, the Encoding and Coherence Process occurs in the same manner as it did with the source text in the Text Representation stage, but this time with the question and answer choices. As before, the vocabulary demand of the question and answer choices is hypothesized to affect the Encoding and Coherence Process. This is followed by Text Mapping, which involves the reader attempting

to locate the information asked by the question in the passage so that s/he can make response decisions. This step is hypothesized to be affected by the amount of text required to be read in order to answer the question as well as the location and format of the key information—whether the information asked is spread out across the passage and whether the information appears verbatim or is paraphrased in the passage. The final step in the Response Decision is Evaluate Truth Status, which involves the reader confirming the correct answer and falsifying distractors. Naturally, the difficulty of this last step would be affected by whether the source passage confirms the correct choice and falsifies distractors (i.e., it is easier if the source passage explicitly confirms the correct choice while falsifying many of the distractors).

To validate this model, Embretson and Wetzel investigated the effects of the passage and item features on the estimated difficulty of 75 RC items in the Armed Services Vocational Aptitude Battery (ASVAB), using the linear logistic latent test model (LLTM, Fischer, 1973). As I discuss more fully in the last section of this chapter as well as the next chapter, LLTM attempts to *explain* the differences between items in terms of the effects of item properties on the probability of correct item responses, rather than simply *describing* the location of items and students on a common scale (De Boeck & Wilson, 2004). The second panel from the top in Figure 3 shows the passage or item features that had statistically significant effects on the difficulty of the RC items. Embretson and Wetzel showed that the item feature predictors at the Response Decision phase had more influence on item difficulty than the passage feature predictors at the Text Representation stage. Based on their findings, Embretson and Wetzel suggested that two separate abilities were involved in ASVAB: verbal ability (or lexical knowledge) required for text processing, and reasoning ability for selecting the correct answer choice, and it was the latter that was emphasized in ASVAB.

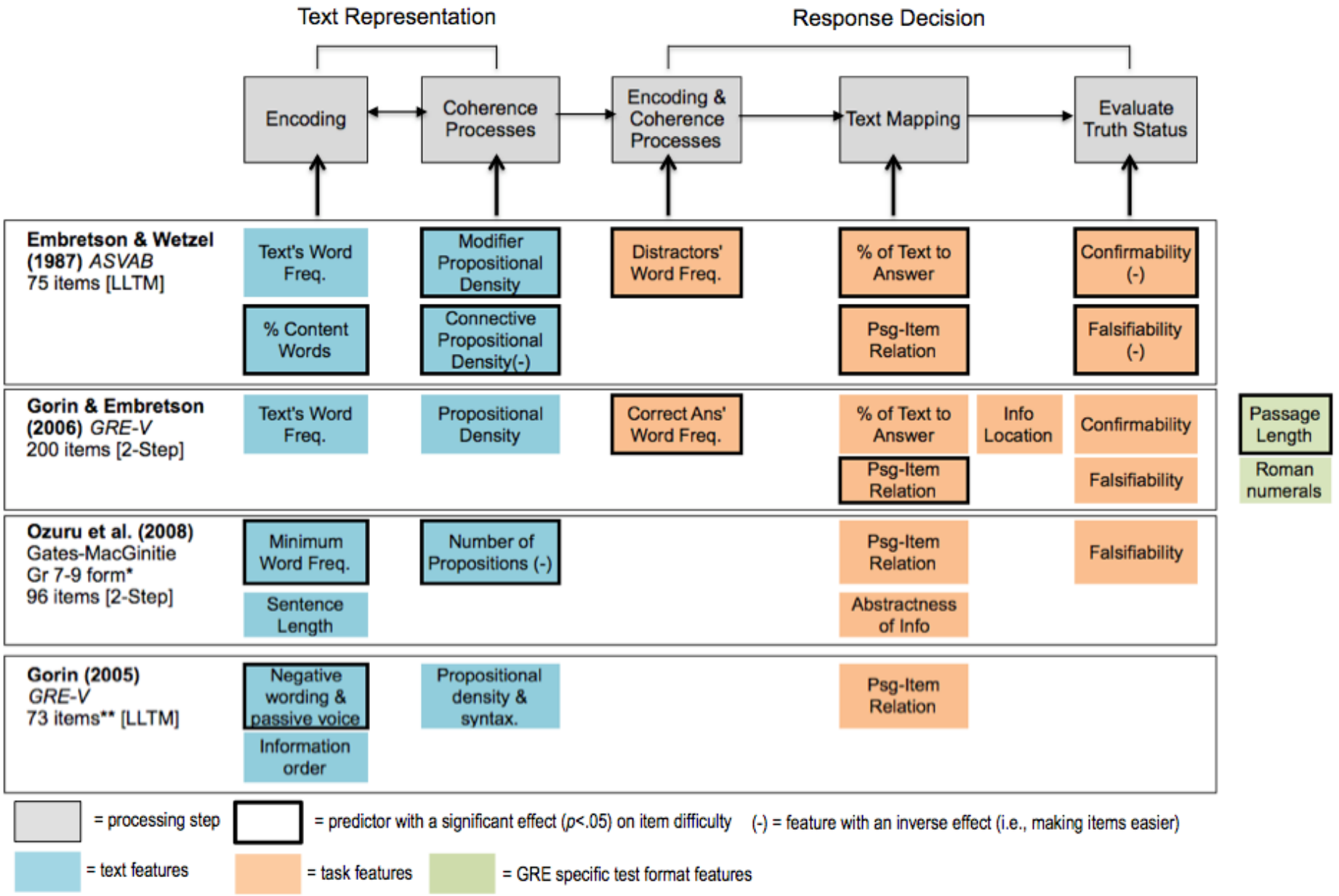

Figure 3. Embretson and Wetzel's information model for multiple-choice paragraph comprehension items and subsequent studies

Three subsequent studies used Embretson and Wetzel's cognitive processing model to empirically identify the item and passage features that underlie student RC performance (Gorin, 2005; Gorin & Embretson, 2006; Ozuru, Rowe, O'Reilly, & McNamara, 2008). The bottom three panels in Figure 3 list the Text Representation and Response Decision variables that were reported to have statistically significant effects on item difficulty from each of these studies. The first of these studies, Gorin & Embretson (2006), examined 200 RC items from 37 passages in the verbal section of the Graduate Record Examination (GRE-Verbal), using a two-step analytical process: first item difficulty parameters were obtained from an IRT model, followed by regressing the item difficulty parameters on different sets of passage and item feature predictors. Similar to the findings in Drum et al. (1981) and Embretson and Wetzel (1987), Gorin and Embretson found that it was item features, rather than the passage features, that affected the item difficulty. Specifically, four item features had statistically significant effects on comprehension performance. Of these four, two represented the Decision Process stage, namely the vocabulary demand of the correct answer choice and the amount of lexical overlap between the passage and answer choices. The remaining two significant predictors were GRE specific item features (i.e., paragraph length and use of roman numerals). None of the text feature predictors that represent the Text Representation stage had a statistically significant impact on student performance.

Three possible explanations were offered for the lack of passage feature effects (Gorin & Embretson, 2006): First, the RC construct measured by the GRE-Verbal items may focus more on student abilities related to Decision Process (i.e., higher-level verbal reasoning ability) rather than passage based comprehension abilities to construct a text representation. The second possibility is that the text feature variables used in the study, namely modifier propositional density, predicate propositional density, and text content word frequency, may not be sensitive enough to account for the variance in IRT-based item difficulty. Thirdly, there was not enough variability in the GRE passages to capture the passages' effect on item difficulty.

The second study that applied the Embretson and Wetzel's information processing model is Ozuru, Rowe, O'Reilly, & McNamara (2008). In analyzing RC items from the Gates-MacGinitie Reading Tests-Reading Comprehension (GMRT-RC), the study arrived at a somewhat different conclusion from the other studies reviewed above. In this study, the published item difficulties (i.e., the proportion of examinees in the norming sample who correctly answered an item) were regressed on several passage and item predictors, some of which were suggested by the Embretson and Wentzel model. Two separate hierarchical linear regression models were performed, one for the 7^{th}-9^{th} grade test form and the other for the 10^{th}-12^{th} grade test form, taking into account that items were nested in passages. The 7^{th}- 9^{th} grade level analysis indicated that item difficulty was primarily affected by text features, as statistically significant effects were found for two of the three passage feature predictors examined, namely, minimum word frequency in the passage and propositional density. In contrast, only one task feature, falsifiability, had a marginal effect that was approaching to significance (p=.09). However, these patterns were not replicated with the analysis of the 10^{th}-12^{th} grade data. In fact, the study found no main effect of passage and item features in the 10^{th}-12^{th} grade analysis.

Gorin (2005) also used Embretson and Wetzel's model, however, unlike the other studies that were all correlational in nature, Gorin experimentally manipulated passage and item features of the released GRE-Verbal RC items. Of the four types of stimulus manipulations investigated, only one passage feature variable, namely the use of passive voice and negative wording in the passage, was found to have a statistically significant impact on item difficulty. Gorin concluded

that the lack of statistical significance might be due to the constrained lengths of GRE passages (capped either at 150 words or 450 words). In such short passages, propositional density cannot vary very much among passages. Heterogeneity of processing processes and strategies among participants was also pointed out as another possible reason why hypothesized passage and item features did not yield expected impact on item difficulty.

Going beyond the item difficulty modeling research that focuses solely on the passage and item features as predictors, a recent study (Kulesz, Francis, Barnes, & Fletcher, 2016) investigated simultaneous effects of the reader, the passage (e.g., word frequency, text cohesion), and the item characteristics to explain individual differences in secondary school students' RC performance on the GMRT subtests. In applying explanatory item response models, the study found that vocabulary and background knowledge were the most robust reader characteristics explaining the variance in the RC ability, while passage genre (i.e., expository vs. narrative) explained the most variance in item difficulty. Additionally, several significant two-way interactions were found among the reader, the passage, and the item factors. For example, the reader's background knowledge had differentially larger effect on low-cohesion passages as compared to passages with high cohesion. These interaction effects substantially reduced variability in item difficulty but not much on the reader's RC ability. Further, the interaction effects were found to be much smaller relative to the large main effects uncovered in the study.

To summarize, the findings from the prior research that modeled item difficulty as a function of passage and item features are mixed. Drum et al. (1981), Embretson and Wetzel (1987), and Gorin & Embretson (2006) indicated that it is mainly the variables reflecting the Response Decision stage where the reader interacts not only with the passage but also with the item and answer choices, that affect the item difficulty. In contrast, Ozuru et al (2008) and Gorin (2005) suggest that it is passage features, such as the passages' vocabulary demand, propositional density, and negative wording/passive voice, that impact student RC performance. More recently, using a more complex model for the reader, Kulesz et al. (2016) found that the text genre was the only test design variable that had a significant effect on item difficulty across the two GMRT test forms.

One of the weaknesses of this line of inquiry is that researchers tended to include only a narrow range of passages in terms of their complexity. These passages are designed for readers at a particular age range (e.g., GRE and ASVAB are for matured readers, while the GMRT forms analyzed are for middle and high school grades). This has likely limited the variability of the passage features, which, in turn, affects their explanatory power (Gorin, 2005, Gorin and Embretson 2006). Relatedly, because the prior research focused on a relatively narrow range of readers, differential impacts of passage and item features by readers' developmental levels were not fully investigated. Indeed, Drum et al. (1981), Ozuru at al. (2008), and Kulesz et al. (2016) did involve students from several grade spans, and Drum et al. (1981), in particular, attempted to discern differential effects of passage and item features by student developmental levels. However their method was descriptive in nature, and did not formally model the interactions between the stimuli features and student grade levels. Ozuru et al. (2008) and Kulesz et al. (2016) both used the GMRT 7^{th} - 9^{th} grade and the 10^{th} - 12^{th} grade forms, but without common items linking the two forms, they could not combine the response data from the two forms to investigate the effects of the passages and the items for a wider grade range of readers.

In terms of statistical models, Embretson and Wetzel (1987) and Gorin (2006) used the linear logistic test model (LLTM; Fischer, 1973), which is considered as a more appropriate statistical model than the two-step approach used in Gorin and Embretson (2006) and Ozuru et

al. (2008). The two-step approach consists of estimating item difficulty estimates first, followed by regressing the estimates from the first step on item and passage predictors). The LLTM is superior because it incorporates external test design variables into the measurement model and hence provides better estimates (i.e., with smaller standard errors) for both item feature difficulty and student ability. Kulesz et al. (2016) used a somewhat more complete model, LLTM+e, which incorporates random variation in item difficulties. LLTM+e overcomes the strict and unrealistic assumption of the LLTM that posits item features in the model account for all variation in the item difficulties (i.e., the LLTM does not allow random variation in item difficulties. Mathematical explication of this difference is offered in the last section of this chapter).

However, to date, few K-12 studies have been conducted applying LLTM or LLTM+e to test data that involves a wide spectrum of readers and passages. Thus, the current study is an attempt to fill this gap by applying explanatory item response models (for details, see below) to testing data from an operational online assessment, which included a wide range of passages and readers.

Quantitative Analytical Tools of Text Complexity

To identify a text's readability or complexity, hundreds of quantitative tools have been developed since the early 1920s (Klare, 1984). At least three generations of quantitative text analyzers can be identified in the literature.

First-generation tools. The first-generation tools typically rely on word difficulty and sentence difficulty in determining a text's readability, with the calculation done by hand or by reference to some handy conversion tables. Examples of such tools include the Flesch-Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975) and Fry Readability Graph (Fry, 1977). Of the first-generation formulas, the Flesch-Kincaid is the most prominent in use today (as part of most word-processing programs).

The Flesch-Kincaid formula was originally developed for Navy use and essentially is a multiple regression equation:

$$\text{Readability of Text (in grade level)} = 0.39 * \text{ASL} + 11.8 * \text{ASW} - 15.59, \quad (1)$$

where ASL represents average sentence length and ASW represents average number of syllables per word. As Equation 1 shows, 0.39 and 11.8 are the weights given to each of the two predictors. The readability of passage— the outcome variable— is expressed in the U.S. grade levels (e.g., a score of 10 means Grade 10 level readability). The Flesch-Kincaid Grade Level robustly predicts reading time of the passage in question. Longer words are more infrequent, thus the average number of syllables per word may account for word knowledge as well as prior knowledge about the topic. The average sentence length, in contrast, may account for the reader's working memory and other cognitive resources to handle more complex syntactic construction.

Second-generation tools. The second generation of tools utilized an increasing amount of computer power to automate analysis. The automated analysis also enabled scholars to consult large corpora of text (rather than sampling its sections) in constructing new formulas. Interestingly, however, word and sentence factors— the same two foci of the first-generation tools—dominated these new analyses. The Lexile Framework for Reading (hereafter Lexile, Stenner, Burdick, Sanford, & Burdick, 2006) and Degrees of Reading Power (DRP, Koslin, Zeno, & Koslin, 1987) are examples of second-generation tools.

Of digitized text analysis systems, the most widely used is Lexile. Its influence has expanded dramatically in recent years, most likely because of its use by the Common Core State Standards (CCSS) in defining the appropriate text levels at different grade bands (Appendix A of

the Standards; National Governors Association [NGA] & Council of Chief State School Officers [CCSSO], 2010). Notably, the developers of Lexile claim that it is not a readability formula (Smith, Stenner, Horabin, & Smith, 1989). Rather, it is an attempt by Stenner and his colleagues to place readers (or their reading ability) and books (or their readability) on a common interval scale with a standardized unit called Lexile (L for short) to facilitate the "best match" between the reader and the text.

Operationally, Lexile uses Cloze "items" that are short passages (each with 125-140 words) ending with a sentence missing a word. A reader is asked to "close" this ending sentence by selecting a word among four answer choices, all of which would work grammatically but only one makes sense semantically. Item writers (or a machine) can create multiple ending sentences for a given passage as shown in Figure 4. Three possible ending sentences are shown in the second column, each with a blank to be filled with a word from four choices.

Passage (430L)	Ensemble cloze items	Answer choices				Observed Difficulty
She disappeared through the trees. "Fine with me," I thought angrily. It would be fine with me if I never saw her again.	(1) I am glad she is _____.	gone	first	best	sitting	269L
	(2) I was _____.	upset	happy	polite	hungry	632L
	(3) I _____ her.	dislike	forgot	told	chased	704L

Figure 4. Lexile's sample ensemble Cloze items

According to Stenner et al. (2006), these three items constitute an ensemble and each item has an associated observed difficulty (see the right most column in the figure). And it is the ensemble's mean difficulty, calculated by averaging over the distribution of the item difficulties, that can be predicted by two traditional textual factors as shown in the following equation:

$$\text{Readability of text (in logit}^2\text{)} = (9.822*\text{LMSL}) - (2.146*\text{MLWF}) - \text{constant}, \quad (2)$$

where LMSL is log of the mean sentence length and MLWF is the mean of the log word frequencies. Notice that the equation is essentially a readability formula with the two traditional textual components: a semantic component and a syntactic component. Stenner and colleagues call this a "specification equation" which, according to them, embodies substantive reading theory and provides the meaning of the construct being measured, including the student's ability to handle the semantic and syntactic demands of a passage (Stenner, Smith, & Burdick, 1983; Stenner, Fisher, Stone, & Burdick, 2013).

Essentially, the passage difficulty (430L in the example in Figure 4), is the amount of reading comprehension ability that the reader needs in order to achieve a 75% success rate across the many possible ensembles of items for the passage. By using these ensembles' mean difficulties, rather than individual item difficulties, Stenner et al (2006) reported that uncertainty associated with the Lexile measures is greatly reduced. Additionally, the authors claim that the averaging across the multiple items in each ensemble removed any particularities associated with the items (e.g., item writers' effect). They further argue that the specification equation enables the creation of strictly parallel items and that their difficulties can be obtained by the theory rather than particular students' response data. This assures that the measurement of the reader

[2] A logit is a unit of measurement that represents logarithm distance between the reader's ability and the text's difficulty, and one logit equals to 180L. See Stenner (1996) for details about rescaling text's difficulty on the logit scale to the Lexile scale.

and the text to be independent of particular items, particular people, and particular contexts, thereby achieving what Stenner and colleagues call "general objectivity" (Stenner et al., 2013).

As this description shows, Stenner and colleagues emphasize that Lexile is based on a substantive theory as embodied in the specification equation. They argue that the theory explains the black box of how the construct being measured (e.g., reading comprehension ability or readability) causes observed variance in response behavior (e.g., a raw count or proportion of correct responses to Cloze items). Specifically, for the semantic demand, Stenner (1996) refers to "the exposure theory," citing Bormuth (1966), Klare (1963), Miller & Gildea (1987), and Stenner et al. (1983), which postulates that words that are used frequently in writing or orally are more likely to be part of one's receptive vocabulary and thus their meaning can be easily recalled. Thus, passages with more familiar words are easier to process than passages with rarer words.

For the syntactic demand, Stenner (1996) once admitted that sentence length is only a proxy, citing the work by Chall (1988) and Davidson and Kantor (1982), which are critical of using readability formulas to modify texts. That is to say, sentence length is not an underlying a causal factor of reading difficulty, and shortening a sentence may increase text's difficulty rather than decrease it. However, more recently, Stenner et al. (2014) has made an argument that is rather contradictory to this admission; they said that the two features included in the specification equation, namely the log word frequency and mean sentence length, constitute the measurement mechanism, which is causally responsible for transmitting the variation in observed comprehension rate at the ensemble level.

In spite of its wide spread use, it is unclear whether the reading research community has accepted Lexile to be solidly grounded in a reading theory. In 2001, a five-member expert panel was convened to examine the theoretical underpinning of the Lexile and its construct validity. Larson (2001), one of the experts, pointed out that Lexile is broadly aligned with widely held "bottom-up" views of language processing, which posit that linguistic input is segmented into sentences, phrases, and words. Word meanings are looked up, and meanings of larger segments are computed from the word meanings. However, this view ignores other important aspects of language processing such as context-dependent elements such as pronouns and deictic motion verbs (e.g., me, come), which are hard to fully comprehend without additional contextual information about the identity and the location of the speaker or the referent. Consider, for example, "He will bring that tool tomorrow." Lexical look up of words "he" "that" and "tool" does not help comprehension of the sentence very much. Rather it depends on the reader/listener's extra-linguistic knowledge, such as who is referred to as "he" and what is "that tool".

Similarly, Kamil (2001), another expert, noted that Lexile may be only adequate if one views reading as a transmission process, where information is transferred from the page to the reader. Pearson & Cervetti (2015) refers to this conceptualization "text-centric", pointing out that this was the dominant theoretical perspective prior to 1965. However, if one takes a more contemporary perspective that sees reading as multidimensional and interactive processes, Lexile falls short as it ignores what the reader brings to the act of reading, including interest, motivation, and background knowledge (Anderson & Davison, 1986; Bruce & Rubin, 1988; Kamil, 2001), and how the reader factors interact with the task and the text factors (RRSG, 2002). Additionally, the Cloze tasks capture only a lower level of comprehension, such as understanding ideas within a sentence (Shanahan, Kamil, & Tobin, 1982), therefore Lexile may

not capture the full range of the reading comprehension construct, especially top-down processes (e.g., inferencing based on prior knowledge) and more global levels of comprehension.

At a more fine-grained level, other two experts on the panel, Adams (2001) and Smith (2001), noted that the word frequency would not fully capture the semantic complexity of the continuous text because of the skewed distribution of words. Specifically, Adams (2001) pointed out that, "75% of the running text was accounted for by just 1,000 of those types and 90% of the text was accounted for by just 5,000 different types" (p.17). Because of this skewness, true frequencies of relatively uncommon words are very difficult to discern but it is these rare words, according to Adams, that matters most in determining the semantic complexity of texts.

In terms of scaling, the Lexile measure typically ranges from below 0 to 2000L, with 200L anchored at the difficulty of first grade basal texts and 1000L at that of a typical passage from an encyclopedia (Stenner et al., 2006). The developers of Lexile claim that the Lexile scale is an interval scale, with one unit having the same meaning across the entire range of the scale (Stenner et al., 2006; Stenner et al, 2013). However, this claim has recently been challenged by several psychometricians (Briggs, 2013; Domingue, 2014; Markus & Borsboom, 2013).

Third-generation tools. Recent years have seen the rise of a third generation of quantitative analytical tools that go beyond the traditional two factors (i.e., word and sentence difficulties), which are grounded in theories of language and discourse processing and text comprehension (e.g., Graesser & McNamara, 2011; W. Kintsch, 1998; Walter Kintsch & Van Dijk, 1978). These theories offer multilevel, multidimensional, and interactive frameworks and are in closer alignment with the RAND heuristic of reading comprehension reviewed earlier. The Coh-Metrix (Graesser, McNamara, & Kulikowich, 2011) and the TextEvaluator (Sheehan, Kostin, Napolitano, & Flor, 2014) are examples of the third-generation tools.

Coh-Metrix. Coh-Metrix is an automated text analysis tool that provides over 100 linguistic indices that cut across, word, sentence, and paragraph/discourse levels (Graesser et al., 2011). Of them, the primary bank of indices examines cohesion. Cohesion refers to the extent to which ideas presented in a text are connected. As such it is a property of the text. In contrast, a closely related term, coherence, refers to the extent to which the reader's mental representations from the text are connected. Thus coherence resides in the mind of the reader and is a product of the reader-text interaction. Further, coherence is affected by cohesive cues embedded in the text (McNamara, Graesser, McCarthy, & Cai, 2014).

Examples of cohesive devices include the use of pronouns and words that are similar in meaning to create referential and semantic overlap across sentences. Connectives such as *because*, *however,* and *before* also provide cohesive ties that make relations among ideas or events explicit, thereby helping the reader develop a more coherent mental representation. Conversely, a break in a certain dimension of cohesion, say temporality as indexed partially by consistency in tense and aspect of main verbs, would pose more demand on the reader to slow down and fill the gap through generating inferences based on prior knowledge, unless the text provides a transitional phrase or other linguistic devices to bridge the gap.

The developers of Coh-Metrix argue that the tool taps into reader's deeper levels of comprehension that goes beyond word and sentence levels because of the explicit focus on both cohesion that cuts within and across sentences and the discourse level features such as text genre. Coh-Metrix also enables one to understand the processing demand of a text for the reader in generating inferences to bridge cohesion gaps. This is a stark difference from the traditional readability formulas, according to McNamara and her colleagues, which focus on the surface

features at the word and sentence levels, hence, are able to capture student's understanding only at those levels.

Another uniqueness of Coh-Metrix lies in its conceptualization of text difficulty as inherently multidimensional. Thus, until recently, Coh-Metrix did not offer a single index of text complexity (McNamara et al., 2014). Instead, over 100 linguistic and discourse measures are combined into eight orthogonal components through principal component analysis (PCA). These components, taken together, accounted for 67.3% of the variability among the criterion texts that were used to develop Coh-Metrix (Graesser, et al., 2011). These components are: (1) narrativity, (2) referential cohesion, (3) syntactic simplicity, (4) word concreteness, (5) deep cohesion, (6) verb cohesion, (7) logical cohesion, and (8) temporal cohesion (short descriptions of these components are provided in the Methods section). Note that the Coh-Metrix developers label these components as "easability" factors, meaning that texts that are higher on these components are easier to process and comprehend (McNamara, et al., 2014).

Graesser and his colleagues argue that having multiple dimensions to represent text variation is important because some dimensions may compensate for each other. For example, when describing unfamiliar scientific concepts to the reader, a textbook author may use more cohesion devices and simpler grammatical construction (e.g., less negation) to make the text more comprehensible (Graesser et al., 2011). They also argue that the multidimensional characterization of text complexity helps generate guidance for diagnosing and responding to students' strengths and weaknesses, thereby facilitating a better match between the reader and the text. Further, Coh-Metrix multiple indices are designed to help teachers learn about potential sources of challenges in texts, which in turn helps them select texts and plan for instructional tasks that require students to recognize the challenges and offer scaffolding to overcome them (McNamara, et al., 2014).

Lastly, Coh-Metrix measures of cohesion hold promise in providing feedback to writing and making texts more comprehensible. In fact, a review of 12 experimental studies investigating the effect of cohesion manipulation on multi-paragraph full texts found robust effects of cohesion on comprehension across populations (students in grades 3 through college), text genres, manipulation techniques, and comprehension measures although cognitively less demanding comprehension measures did not yield the same level of effects as more challenging measures that require inferencing and keyword sorting (McNamara, Louwerse, McCarthy, & Graesser, 2010). Notably, one of the studies (Britton & Gülgöz, 1991) demonstrated that students' comprehension improved more when they read texts whose cohesion had been improved rather than the texts that were revised according to a traditional two-factor readability formulas (i.e., shortening sentence length and using more familiar words). One of the intriguing aspects about cohesion improvement is that it typically leads to an increase in text's readability as estimated by a traditional two-factor measure. This is so because the revision typically involves making sentences longer by using connectives (e.g., *because*, *therefore*) to combine clauses and sentences, and to make their relationship clearer. The revision could also involve adding more information to the text so that readers do not have to make knowledge-based inferences (see example 2 in Figure 5). Further, the cohesion revision tends to increase the use of rare words as it replaces pronouns with specific referents to increase referential overlap across sentences. For example, "they" is replaced with "the bombing attacks" in the second example of cohesion revision in Figure 5.

Original, Low-Cohesion Text	Modified, High-Cohesion Text
Example 1	
Smoking was forbidden. The store had inflammables.	Smoking was forbidden *because* the store had inflammables.
Example 2	
Most members of the Johnson administration believed bombing attacks would accomplish several things. They would demonstrate clearly and forcefully the United State's resolve to halt communist aggression and to support a free Vietnam.	Most of *both civilian and military* members of the Johnson administration believed bombing attacks would accomplish several things. *The bombing attacks* would demonstrate clearly and forcefully the United State's resolve to halt *communist North Vietnam's aggression* and to support a free *South Vietnam*.

Figure 5. Examples of cohesion revision. Source: McNamara et al. (2014)

Coh-Metrix is based on decades of research on the effects of cohesion on discourse processing and comprehension (for a summary of this research, see McNamara et al., 2010; 2014). However, its validity evidence is rather limited. The development of text analysis tools typically requires passages with predetermined complexity scores as criterion variables. Cunningham & Mesmer (2014) observed that historically the criterion variables came from (a) an existing readability formula, (b) authors or publishers of the passages or based on traditional use of passages in schools, (c) teachers' or other experts' judgments, or (d) students' reading comprehension performance. Of them, (d) student performance is the most direct measure of passage difficulty, thus Cunningham and Mesmer argues that it should be the gold standard. Further, they argue that the most genuine standard (i.e., even more "gold" standard) is students' growth in reading comprehension after having given a particular set of texts selected or designed by the text complexity tool to be validated. This argument makes sense from the consequential validity perspective (Messick, 1989) in the era of the Common Core State Standards, which prescribe a particular range of text complexity to students at a particular grade level band. In other words, CCSS has created something like a marketplace for the more prominent use for text complexity tools than ever before.

Coh-Metrix eight PCA components were calibrated with 3,900 one-paragraph academic texts sampled from the Touchstone Applied Science Associates (TASA) corpus. Each TASA passage is indexed with a grade level determination by an existing readability measure called Degrees of Reading Power (DRP; Koslin, Zeno, & Koslin, 1987). Like the Flesch-Kincaid and the Lexile, the DRP relies on three measurable text features at the word and sentence levels: word familiarity, word length, and sentence length. Although the developers of Coh-Metrix justify the use of the TASA corpus by characterizing it as one of the most representative of texts that a typical senior in high school would have encountered throughout their K-12 schooling, its appropriateness is questionable, for at least two reasons: (a) the TASA passages were artificially truncated and (b) their difficulty (the criterion variable) was determined by DRP— a readability formula that relies on the traditional word and sentence factors.

As Sheehan et al (2014) has aptly pointed out, the TASA passages were artificially truncated at about 300 words (without paragraph marks) for creating a word-frequency index. Applying this truncation strategy to their own corpus (called the TextEvaluator corpus, more on

this in the next section), Sheehan et al. (2014) showed that the number of sentences and paragraphs in truncated passages, on average, is half of their original length, which raises a concern that the TASA corpus may not fully represent the difficulty associated with making connections across the complete arc of sentences and paragraphs in naturally occurring passages. This is an unfortunate limitation given that Coh-Metrix's primary goal was to incorporate multiple levels of textual features in estimating texts' complexity, especially at the discourse level, which necessarily involves processing and comprehension challenges that stretch across sentences and paragraphs.

Another drawback of TASA passages is that its difficulty was determined by DRP. With the DRP difficulty index as the criterion variable, Coh-Metrix PCA measures may have been reduced to detect only the surface level comprehension that can be captured by the traditional word and sentence level features. Ironically, the explicit goal of Coh-Metrix was to overcome the limitations of the existing readability formulas, and yet Coh-Metrix eight summative measures (the PCA components) were calibrated to best predict the difficulty of the artificially short TASA passages as determined by DRP, which relies on the word and sentence level predictors that the Coh-Metrix developers criticized as inadequate. Moreover, to date, the Coh-Metrix measures have been validated mostly with criterion passage ratings that are based on human (either novices or experts) judgment (e.g., McNamara et al., 2010) and few studies seem to have been conducted with corpora scaled on direct measures of student performance.

An exception might be a comparative study of several text analysis tools by Nelson, Perfetti, Liben, & Liben (2012), which included the Lexile, Coh-Metrix and the TextEvaluator which are reviewed in this chapter. As part of the study, pairwise rank-order correlations (Spearman's rho) were examined between one of the five Coh-Metrix PCA-based measures, on the one hand, and criterion passages' difficulty as determined by averaging Rasch-based item difficulty at the passage level, on the other. The criterion variables were made available from the publishers of the following three assessments: the GMRT-RC, the Stanford Standardized Assessment, Ninth Edition (SAT-9), and MetaMetrics Oasis. Interestingly, among the 15 correlation coefficients examined (five Coh-Metrix PCA components × the three assessments), the strongest relationship (rho ≈ -.82[3]) was found between Coh-Metrix syntactic simplicity and MetaMatrix Oasis passages' difficulty, which was entirely determined by the type of the Cloze items reviewed earlier. The negative coefficient means that texts with higher syntactic simplicity scores (i.e., with familiar and simpler grammatical structures) are easier as determined by student performance on the Cloze tasks. The second largest correlation (rho ≈ - .67) was also found between MetaMatrix Oasis passages' difficulty and Coh-Metrix narrativity measure, indicating that easier texts are the ones with higher narrativity scores (this means that easier passages contain a higher proportion of more-story like features such as familiar words, pronouns, and higher ratios of verbs-to-nouns). The other two assessments, the GMRT-RC and SAT-9 have more varied item types such as questions that require literal recall vs. inferences, albeit all are in the multiple-choice format. These two assessments also correlated with the two Coh-Metrix measures (i.e., syntactic simplicity and narrativity), but the magnitude of the association was substantially lower (-.57 ≤ rho ≤ -.35) than those found with the MetaMatrix Oasis passages.

To sum up, Coh-Metrix is one of the third generation text complexity tools that incorporate multiple levels of factors that affect text complexity, with an explicit focus on

[3] The coefficient was read from a graph (Figure 5.4–3) in Nelson et al. (2012, p.45) therefore the number may not be exact.

cohesion. While it has solid theoretical basis, Coh-Metrix eight "easability" factors might be limited in their predictive power as they were calibrated with relatively short passages whose difficulty had been determined by the traditional word and sentence level features.

TextEvaluator. Like Coh-Metrix, the TextEvaluator system is grounded in theories that view reading as an active process through which readers seek to build coherent mental representations of the information presented in the text (e.g., Alderson, 2000; Gernsbacher, 1990; Just & Carpenter, 1987; W Kintsch, 1988; RRSG, 2002). As such, TextEvaluator is designed to reflect not just the lower level processes related to the understanding of words and sentences, but also the higher level processes such inferring connections across sentences using textual cohesive clues (e.g., repeated content words and explicit connectives) as well as prior knowledge and experience, in an effort to develop more complete and integrated mental representations of text—what Kintch (1988) called the situation model.

To achieve this goal, the TextEvaluator establishes text's complexity on eight dimensions that cut across word, sentence, and discourse levels: (a) academic vocabulary, (b) syntactic complexity, (c) word concreteness, (d) word unfamiliarity, (e) interactive/conversational style, (f) degree of narrativity, (g) lexical cohesion, and (h) argumentation (Sheehan et al., 2014). Like Coh-Metrix, these eight dimensions were derived from principal component analysis, representing patterns of co-occurrence of many correlated linguistic features in the text (Sheehan et al., 2014). In the development of TextEvaluator, 43 text features were reduced into the eight principal components that, in concert, accounted for over 60% of variation in text complexity across a wide range of passages.

One of the unique features of the TextEvaluator is its explicit attention to a genre bias in predicting text complexity. Typically, difficulty of informational text is overestimated because of the repetition of rare content words while that of narrative text is underestimated due to the prevalence of short dialogues. Sheehan (2016) observed such bias with the two popular readability formulas: the Flesh-Kincaid Grade Level and the Lexile. TextEvaluator overcomes this bias by using three distinct prediction models optimized to three text types: narrative, informational, and mixed (Sheehan et al., 2014; Sheehan, 2016).

For the validity of the TextEvaluator, Sheehan and her colleagues have shown that not only the word- and sentence-level PCA components, but also most of those related to the discourse-level and the use of prior knowledge, uniquely contributed to the prediction of the complexity of criterion passages. And this was true for a set of informational texts as well as for a set of literary texts. Further, it has been shown that the TextEvaluator's overall text complexity scores are highly correlated (r =.72-.91) with the grade levels of criterion passages in the TextEvaluator corpus, which included texts from a variety of sources, ranging from high-stakes state, national, and college admission assessments (e.g., NAEP, the SAT®) to CCSS' exemplar passages. One caveat, however, is that the TextEvaluator's calibration and validation relied on the grade-level determination made solely by human judges, rather than by student performance. What this means is that all passages from fourth-grade reading assessments, for example, are classified as grade 4 in terms of text complexity. The TextEvaluator corpus does include passages graded using more fine-grained criteria, such as exemplar passages from the Appendix B of the CCSS (CCSS Initiative, 2010) and from Chall, Bissex, Conard, & Harris-Sharples (1996). Further, Chall et al.'s exemplar passages were partially validated with students' Cloze comprehension scores. However, these exemplar passages constitute only about 18 percent of overall the passages included in the TextEvaluator corpus used to develop, train, and validate TextEvaluator.

To sum up, TextEvaluator is another example of the third generation text analyzers grounded in contemporary theories that view reading as active and complex processes. Like Coh-Metrix, TextEvaluator incorporates word, sentence, and discourse levels of factors that affect text complexity. However, its predictive power may be limited because it has been trained only with rather coarse difficulty ratings made by human judges. Sheehan et al. (2014) explicitly stated that the goal of TextEvaluator is to develop a common scale for both texts and reader. However, to date, the measurement of reader ability doesn't appear to have gained much traction in the TextEvaluator system.

Figure 6 summarizes the four analytical tools of text complexity reviewed and their linguistic and text factors that are taken into consideration in deriving text complexity scores. All but Coh-Metrix provides scores for overall text complexity, while all but Flesch-Kincaid Grade Level provide subcomponent scores (i.e., the scores for the linguistic/text variables listed in the figure).[4]

	Analytical Tools (developer)	Unit	Linguistic/Text Variables		
			Word Level	Sentence Level	Discourse/ Text Level
Traditional Two-Factor model	**Flesch-Kincaid Grade Level**	Grade level	• Word length	• Sentence length	
	Lexile	Lexile	• Mean log word frequency	• Sentence length	
Newer-Multi-Level-Multi-Factor Model	**Coh-Metrix**	Z-score 1-100	• Narrativity^{+} • Syntactic simplicity^{+} • Word concreteness^{+} • Verb cohesion^{+} • Logical cohesion^{+}	• Narrativity^{+} • Syntactic simplicity^{+} • Deep cohesion^{+}	• Narrativity^{+} • Referential cohesion^{+} • Syntactic simplicity^{+} • Deep cohesion^{+} • Verb cohesion^{+} • Temporality^{+}
	TextEvaluator	1-100	• Word unfamiliarity^{+} • Word concreteness^{+} • Academic vocablary^{+}	• Syntactic complexity^{+}	• Lexical cohesion^{+} • Interactive style^{+} • Narrativity^{+} • Argumentation^{+}

Figure 6. Quantitative analytical tools of text complexity used in the study

Note. $^{+}$ A component derived from multiple variables based on principal component analyses.

As noted earlier, Nelson et al. (2012) conducted a comparative study of several measures of text complexity as part of the CCSS effort. The study established relative efficacy of all three generations of the text complexity measures with rank-order correlations, highlighting somewhat higher coefficients of the third generation tools such as Coh-Metrix and TextEvaluator, especially with texts for older students. However, as suggested earlier in this chapter, these newer tools may be limited by their validation methods, mainly because they have been calibrated with corpora scaled by measures of human judgment rather than direct measures of

[4] Coh-Metrix provides Flesch-Kincaid's Grade Level scores (for overall complexity) as well as scores for word length and sentence length as part of its text analysis.

student reading performance on tasks that require deeper level of comprehension (e.g., making connections across sentence boundaries, and activating prior knowledge to construct coherent and integrated mental models). Even with this caveat, the role that the text analysis tools play in the CCSS policy cannot be overlooked. In fact, the Nelson et al. study resulted in a revised set of CCSS' recommendations regarding the ranges of text complexity to guide teachers in selecting texts appropriate for students in Grades 2-12 (Coleman & Pimentel, 2012). Despite the prominence of these text analysis tools in the CCSS documents, however, these tools have not been fully incorporated into the item difficulty modeling research reviewed above. Thus the current study took advantage of these analytical tools to obtain text features listed as the linguistic and text variables in Figure 6, and investigated the extent to which they predicted the difficulty of RC items.

Explanatory Item Response Models

A standard application of item response models yields measures of items and persons. In the context of the one-parameter logistic (Rasch) model, a person measure captures a latent variable, which could be an ability, trait, or an attitude, while an item measure captures item difficulty. These measures are descriptive in that they inform the person and the item's standing on a common scale. The descriptive approach can be complimented by an explanatory approach (DeBoeck & Wilson, 2004), which seeks to model and explain the probability of a certain response (e.g., answering correctly in the case of binary scored items in the achievement context), using external variables such as characteristics of items (e.g., item type that represents underlying cognitive processes) and/or characteristics of persons (e.g., gender, age, treatment status).

Typically, in the test data, more than one observations are made for each person, and observations are made for more than one person. Figure 7 illustrates a hypothetical data matrix from two people responding to three items (Item 1, Item 2, and Item 3). The matrix is in the wide format, with one row per person and one column per item. It shows that person 1 correctly responded to the first item and incorrectly responded to the remaining two items, with a total score of 1 (out of 3). The last row summarizes the items' summative scores in terms of the proportion of people answering correctly (also known as p-value in classical item analysis).

	Item 1	Item 2	Item 3	*Total scores*
Person 1	1	0	0	*1*
Person 2	1	0	1	*2*
Proportion correct	*1*	*0*	*0.5*	

Figure 7. Illustration of test data in wide data form

The same response data can be represented in the long-data format, along with three item features and one person characteristic as shown in Figure 8. Each row in the figure represents a distinct observation Y_{ip} (1 for a correct response and 0 for a incorrect response), with the first two columns indexing persons (p) and items (i). The next three columns indicate the three item features: the first item feature X_0 shows that all items have a value of 1, serving as a nominalization constant comparable to the intercept in a regression model; X_1 indicates whether items require literal recall, and X_2 indicates whether items call for inferences. As can be seen in the figure, these item features vary across items but stay constant among persons. The last

column, person characteristic Z_0, shows persons' vocabulary score in the grade level unit. As such, it varies across people but remains constant within persons.

person	item	response	item features (X_k)			person characteristic (Z_j)
p	i	Y_{ip}	X_0	X_1 (literal recall)	X_2 (inference)	Z_0 (vocabulary)
1	1	1	1	1	0	3
1	2	0	1	0	1	3
1	3	0	1	0	1	3
2	1	1	1	1	0	6
2	2	0	1	0	1	6
2	3	1	1	0	1	6

Figure 8. Illustration of test data in long data form with item and person characteristics

This long-format data shows a two-level structure where responses are nested within persons. In EIRM, the response Y_{ip} is the outcome variable of interest to be predicted by the item and person characteristics (X_k and Z_j). However, because Y_{ip} is a categorical dependent variable, it needs to be transformed through a link function so that the expected probability of a correct response can be correctly mapped to a linear combination of item and person predictors. The common link function for EIRM is the logit function although other link functions are also possible (see DeBoeck & Wilson 2004 for details). Taking the log-odds or logit of success probability, the Rasch model, which describes the location of items and persons, can be expressed mathematically as:

$$logit\left[\mathrm{P}\left(Y_{ip}=1\,|\theta_p,\beta_i\right)\right]=\theta_p-\delta_i, \tag{3}$$

where p is a person index (p = 1, . . . , P), i is an item index (i = 1, ... I), θ_p is a unidimensional ability parameter, and δ_i is an item difficulty parameter. Individual differences in θ_p can be explained by person characteristics:

$$\theta_p=\sum_{j=0}^{J}\vartheta_j Z_{pj}+\epsilon_p, \tag{4}$$

so that

$$logit\left[\mathrm{P}\left(Y_{ip}=1\,|\theta_p,\beta_i\right)\right]=\sum_{j=0}^{J}\vartheta_j Z_{pj}+\epsilon_p-\beta_i, \tag{5}$$

where ϑ_j is the fixed regression weight of person characteristic j, Z_{pj} is the value of person p on person characteristic j, and ϵ_p is the remaining person effect after the effect of the person characteristics is accounted for (i.e., residual). ϵ_p is considered as a random effect and is assumed to follow a normal distribution: $\epsilon_p \sim N\left(0,\sigma^2_{\epsilon_p}\right)$. This is a person explanatory model, also known as the latent regression model (Adams, Wilson, & Wu, 1997).

Similarly, differences in item difficulty (δ_i) can be explained by item features:

$$\delta'_i=\sum_{k=0}^{K}\beta_k X_{ki}, \tag{6}$$

so that

$$logit\left[\mathrm{P}\left(Y_{ip}=1\,|\theta_p,\beta_i\right)\right]=\ \theta_p-\sum_{k=0}^{K}\beta_k X_{ki}, \qquad (7)$$

where β_k is the fixed regression weight of item features k, and X_{ki} is the value of item *i* on item features k. Comparing Equation (7) with Equation (3), it is clear that item difficulty δ_i is replaced with a linear combination of item features X_{ki} and its regression coefficients β_k. Notice that Equation (6) does not have an item specific random error term, therefore δ'_i in Equation (6) will not equal δ_i in Equation (3) (thus the prim sign is used for the former). Equation (7) is an item explanatory model, also known as the linear logistic test model (LLTM; Fischer, 1973). If we allow residual variation of item difficulties that follow a normal distribution ($\tau_i,\ \tau_i \sim N(0,\sigma^2{}_\tau)$, the right hand side of equation (7) becomes,

$$logit\left[\mathrm{P}\left(Y_{ip}=1\,|\theta_p,\beta_i\right)\right]=\ \theta_p-\sum_{k=0}^{K}\beta_k X_{ki}+\ \tau_i. \qquad (7)'$$

This model with the random item error term is referred to as LLTM+e and is a more complete model on the item side (Janssen, Schepers, & Peres, 2004). However, the estimation of this model is a computationally demanding as both the person ability (θ_p) and item residuals are treated random (thus it is classified as a cross-random effect model; De Boeck, 2008)).

Lastly, both the person ability (θ_p) and the item difficulty (δ_i) can be explained with person and item characteristics:

$$logit\left[\mathrm{P}\left(Y_{ip}=1\,|\theta_p,\beta_i\right)\right]=\ \sum_{j=0}^{J}\vartheta_j Z_{pj}+\epsilon_p-\sum_{k=0}^{K}\beta_k X_{ki}. \qquad (8)$$

Equation (8) is a doubly-explanatory model, which is also known as the latent regression LLTM (Wilson & DeBoeck, 2004).

The explanatory models in Equations (5), (7), (7)' and (8) are not new models but are part of a larger statistical framework called *generalized linear mixed models* (GLMM, Breslow & Clayton, 1993; McCulloch, Searle, & Neuhaus, 2001).

The hallmark of EIRMs is to accomplish the measurement and explanation at the same time. This is a superior method to a two-step approach more prevalent in the literature, because EIRMs directly incorporate the item and person predictors into the measurement model and thus provide better parameter estimates for both items and persons (Briggs, 2008; Hartig, Frey, Nold, & Klieme, 2012; Mislevy, 1987). The two step approach, in contrast, consists of obtaining estimates of students' ability or item difficulty (as total score-based or the Rasch-model based), followed by regressing the derived estimates on explanatory variables such as student characteristics (e.g., gender, SES, ethnicity, and treatment status) and/or design features of items (e.g., cognitive processes involved in solving the item). Note that the two-step approach uses a single column of a dependent variable (e.g., a total person scores or items' p-value in Figure 7) as an observed variable, which is free of measurement errors. This means that uncertainty associated with the ability or the item difficulty estimates are not taken into consideration in the subsequent analysis. In contrast, EIRMs utilize the item-by-person information matrix, as shown in Figure 8 above, thereby accounting for within-person and between-person differences, measurement errors, and dependency in clustered data.

EIRMs are attractive because of their ability to model individual students interacting with each item, just as standard item response models do. But EIRMs go beyond the measurement

practice of locating individuals and items on the common scale; as the name suggests, EIRMs seek to *explain* item responses in terms of person and item properties that co-vary with observed responses (De Boeck & Wilson, 2004). In other words, the focus of analysis becomes examining relationships between item responses and person and/or item characteristics, rather than describing the location of individual items and persons. Such explanatory capacity enables the researcher to examine substantive theories, such as theories of discourse processing or reading comprehension, as they relate to readers' performance and its underlying factors within an assessment context.

Indeed, an emerging body of research applies ERIMs to item response data in the domain of literacy as a way to address substantive theoretical issues related to phoneme segmentation/decoding accuracy (Bouwmeester, van Rijen, & Sijtsma, 2011; Gilbert, Compton, & Kearns, 2011), letter-sound acquisition (Kim, Petscher, Foorman, & Zhou, 2010), lexical representation (Cho, Gilbert, & Goodwin, 2013), spelling (Kim, Petscher, & Park, 2016); reading comprehension (Kulesz, 2014; Miller, Davis, Gilbert, Cho, Toste, Street & Cutting, 2014; Sonnleitner, 2008) and visual processing skills (Santi, Kulesz, Khalaf, & Francis, 2015). For example, Cho et al. (2013) applied an explanatory, multidimensional multilevel random item response model to examine the dimensionality of middle school students' word knowledge (termed lexical representation, Perfetti, 2007), with special focus on the contribution of students' morphological knowledge. Similarly, Kim et al. (2016) investigated young children's spelling performance through EIRMs, which enabled the modeling of simultaneous effects of word characteristics and child factors. Yet another study by Santi and her colleagues (Santi et al., 2015) applied EIRMs to a longitudinal data from K-2 students and examined changes in their visual processing skills as a function of student and item characteristics, with an emphasis on uncovering the unique contribution of students' developing reading skills to their growth in visual processing skills.

To summarize, the present study capitalizes on the explanatory capacity of EIRMs to investigate how "item features" and a reader characteristic contribute to comprehension performance. *Item features* in this study are broadly conceived to include both the features of assessment passages (e.g., various indicators of text complexity) as well as task features (e.g., cognitive demands of reading comprehension questions). The reader characteristic to be examined is general vocabulary knowledge, which was assessed prior to the assessment of reading comprehension within the same online assessment system. Ideally other reader characteristics such as their background knowledge and working memory should be examined as well. However, such information was not readily available for the study. Of particular interest are the interactions among passage features, task features, and the reader characteristic as depicted by the RAND heuristic of reading comprehension. The next chapter provides the research questions addressed in the study, along with the data and methodology used to answer them.

Chapter 3. Research Questions and Method

This chapter presents research questions and method in five sections. The first section presents two research questions that guided the current study, followed by the second section that describes the ReadingPlus Insight Assessment—an operational placement test associated with an online reading intervention program, from which a response data was obtained. The third section details a design for anchoring and sampling to vertically scale the response matrix that spanned a wide range of passages and students. The forth section describes the psychometric models and variables used in the study. The final section briefly describes the analytic process.

Research Questions

This study investigated the influence of passage and task features on the difficulty of RC items, using explanatory item response models. Specifically, the following research questions were investigated:

1. Which set of text and task features best explain variability in the difficulty of RC items after controlling for general vocabulary knowledge of students?
2. Are any of text feature effects moderated by students' vocabulary knowledge and/or by task characteristics?

With respect to the first research question, the study specifically looked into whether newer models of text complexity that include linguistic predictors that go beyond the word and sentence levels, namely the Coh-Metrix and the TextEvaluator, had more explanatory power than Lexile that relies on the traditional two factors—mean sentence length and mean log word frequency. Also investigated was whether item difficulty is better explained by a set of passage features that affect the Text Representation phase of the Embretson and Wetzel model or by a set of task features that affect the Response Decision phrase. To date, findings from the previous studies are mixed on this topic. As for the second research question, text-reader, text-task, and text-reader-task interactions were investigated to unpack which text features matter, for whom, and for which type of comprehension tasks.

The Assessment

The data came from an operational online adaptive assessment called ReadingPlus InSight Assessment, which was used for placement and benchmarking within an online reading intervention program (www.readingplus.com). The assessment is comprised of two parts: (a) general vocabulary knowledge assessment, and (b) reading comprehension assessment.[5] The reading comprehension component is comprised of testlets, each with one passage and five multiple-choice questions (thus the number of items per passage was constant across all testlets). All assessment passages were informational texts and four types of comprehension questions were asked about the source passage: (a) gap-filling, (b) main idea, (c) text-connecting, and (d) literal recall. As described earlier, this assessment did not allow students to access the source passage while they answered the related comprehension questions.

[5] I had no influence over the design of the ReadingPlus Insight Assessment nor its data collection. Two of my committee members, Prof. David Pearson and Dr. Elfrieda H. Hiebert, who had advised the assessment design, enabled me to access the archival data.

Indonesian shadow puppets are one of the oldest forms of storytelling. Artists use buffalo leather to make the flat, intricate figures of the shadow puppets. Then, puppeteers use sticks to move the figures behind a large white screen. As a result, the puppets come to life as shadows on the screen. Long ago, the puppeteers used oil lamps to cast the shadows of the puppets on the screen, but now they use electric lights.

Shadow puppets are made by highly skilled artists. The artists first trace the figures onto buffalo leather, and then they roll the leather flat with a glass bottle. Next, they cut out, prime, and paint the figures. Lastly, they attach sticks to control the puppets. An artist needs several weeks to complete a puppet because each one requires such intricate craftsmanship.

The art of making shadow puppets has not changed much over the past three centuries, but the figures do vary from region to region. For example, one area may create natural figures such as birds and animals, while another area may create figures based on real people.

Shadow puppet shows commonly take place at religious ceremonies, family celebrations, and public events. They are often accompanied by drum music and may last all night. Because they usually tell a moral story, the shadow puppets take part in a battle between good and evil. Many people think shadow puppet shows are one of the finest examples of the art of storytelling, and the puppets are considered works of art.

Question [question-answer relation] [abstractness of info requested]	**Answer choices**	**[falsifiable]**
1. What is the main idea of this passage? *[bridging]* *[highly abstract]*	a. Shadow puppet shows are an old form of storytelling still used today.	[no]
	b. Shadow puppet shows are an extinct form of storytelling.	[yes[a]]
	c. Artists use buffalo leather to make shadow puppets.	[no]
	d. Puppet shows are a popular form of storytelling in South America.	[no]
2. Which of the following is true about shadow puppets? *[bridging]* *[somewhat concrete]*	a. Shadow Puppet shows occur at religious ceremonies and family celebrations.	[no]
	b. Shadow Puppet shows are always performed in complete silence.	[yes[b]]
	c. Shadow Puppet shows are a new art form in Indonesia.	[yes[a]]
	d. Shadow Puppet shows are always about romance.	[yes[c]]
3. How have shadow puppet shows changed over the years? *[bridging]* *[somewhat concrete]*	a. Puppet shadows are now cast by an electric light.	[no]
	b. Puppet shadows are now cast on a white screen.	[no]
	c. Puppets are now based on real people.	[yes[d]]
	d. Puppets are now controlled by strings.	no]
4. How do Indonesians feel about shadow puppet shows? *[knowledge-base]* *[somewhat abstract]*	a. Shadow puppet shows are an important art.	[no]
	b. Shadow puppet shows are only for special occasions.	[no]
	c. Shadow puppet show figures can be made by anyone.	[yes[e]]
	d. Shadow puppet shows are created for kids.	[no]
5. Why are shadow puppets considered works of art? *[bridging]* *[somewhat abstract]*	a. The puppets require skilled craftsmanship.	[no]
	b. The puppets represent a specific region of the country.	[no]
	c. The puppets tell a story about good and evil.	[no]
	d. The puppets are painted on the hands of the artists.	[yes[f]]

Figure 9. Sample testlet "Shadow Puppet" with task feature codes (in brackets)

Note. Explicit textual information from the source passage that can be used to falsify distractors is listed below.
[a] "the art of making shadow puppets has not changed much over the past three centuries"
[b] "They are often accompanied by drum music" [c] "they usually tell a moral story"
[d] "one area may create natural figures such as birds and animals"
[e] "puppets are made by highly skilled artists"
[f] "puppeteers use sticks to move the figures behind a large white screen"
For details of the task feature codes, see the "Task features affecting the response decision phase" section below.

Figure 9 above shows a sample testlet called "Shadow Puppet" with three sets of codes given for task features (for details about the task feature codes, see the section entitled, "task features affecting the response decision phase" below).

In the ReadingPlus InSight assessment system, the testlets were organized into 12 levels, which were determined by the vocabulary demand of assessment passages as measured by Lexile®Analyzer, an online text analysis tool that provides the overall complexity in Lexile, mean sentence length, and mean log word frequency. According to the assessment developer, the 12 text levels roughly correspond to grades 1-12 and were aligned with the Common Core State Standards. The analytic sample for the current study included four testlets per level with a total of 48 passages and 240 items (recall five questions per passage).

The assessment was computer-adaptive and the majority of students took five testlets during one test administration. The first passage was selected based on students' performance on the vocabulary component of the ReadingPlus InSight assessment, given prior to the comprehension component. The second through fourth testlets were given based on student performance on the previous testlet. The fifth passage was chosen randomly from the bank of testlets. The current study utilized item responses from these randomly-selected fifth testlets because they were administered to a wide range of students in terms of their comprehension ability without the assessment's adaptive logic. Thus, the data provides accurate item difficulty estimates, which were to be predicted with passage and item features. To construct a common vertical scale that covers the wide range of testlets (in terms of text complexity) and of students (in terms of comprehension ability), seven anchor passages were identified. The next section describes the anchor design that I used to assemble the testlets and students.

Anchoring Design and Sampling for Vertical Scaling

If the response data were to be taken only from the fifth randomly given testlets, essentially each student would provide their responses to only one of the 48 testlets of different difficulty. Consequently, the resulting data would have no link for item- as well as person-parameters to be placed on an overall common metric. The best link would be established when all 48 testlets, totaling 240 items, were administered to all students. However, this was not ethically and economically possible (imagine second graders required to take all 48 testlets, which vocabulary demand spanning from grades 1 through grade 12). To overcome this issue, I came up with an anchoring design to assemble a response data matrix for vertical scaling.[6]

Vertical scaling is a process of linking different levels of an assessment, which measure the same construct, onto a common developmental metric (Harris, 2007). This procedure enables the comparison of scores from test forms of systematically different difficulty (Patz & Yao, 2006). When the scaling is done with IRT models, the score comparison is possible for student performance as well as for test items on the common developmental scale (note that the latter is the focus of this study).

Let's consider a simple case where two tests are administered to two groups of students. To analyze these two sets of data together, an anchor is needed to link the two data sets. The anchor can be "common items" that are taken by both groups of students or "common people"—a subset of students who have completed both testlets (Vale, 1986). This can be visualized in matrices of students (in rows) and items (in columns) as shown in Figure 10. In each matrix, a cell represents a potential item response. Figure 10A shows the ideal scenario where all students

[6] I am grateful for the advice that I received from Professor Mark Wilson for this anchoring design.

took all items, therefore all the cells are filled with item responses, which is indicated by light gray shade (note, in the actual matrix, each cell will filled in with a score of either 0 or 1 for a dichotomously scored item.) As can be seen, Figure 10A is a complete matrix and no anchor is needed. In contrast, Figures 10B - 10D have cells that are left blank, indicating missing responses. Figure 10B shows a situation where Test 1 was given to one student group and Test 2 was given to another group. There is no overlap of items or students to link the two sets of response data in the matrix.

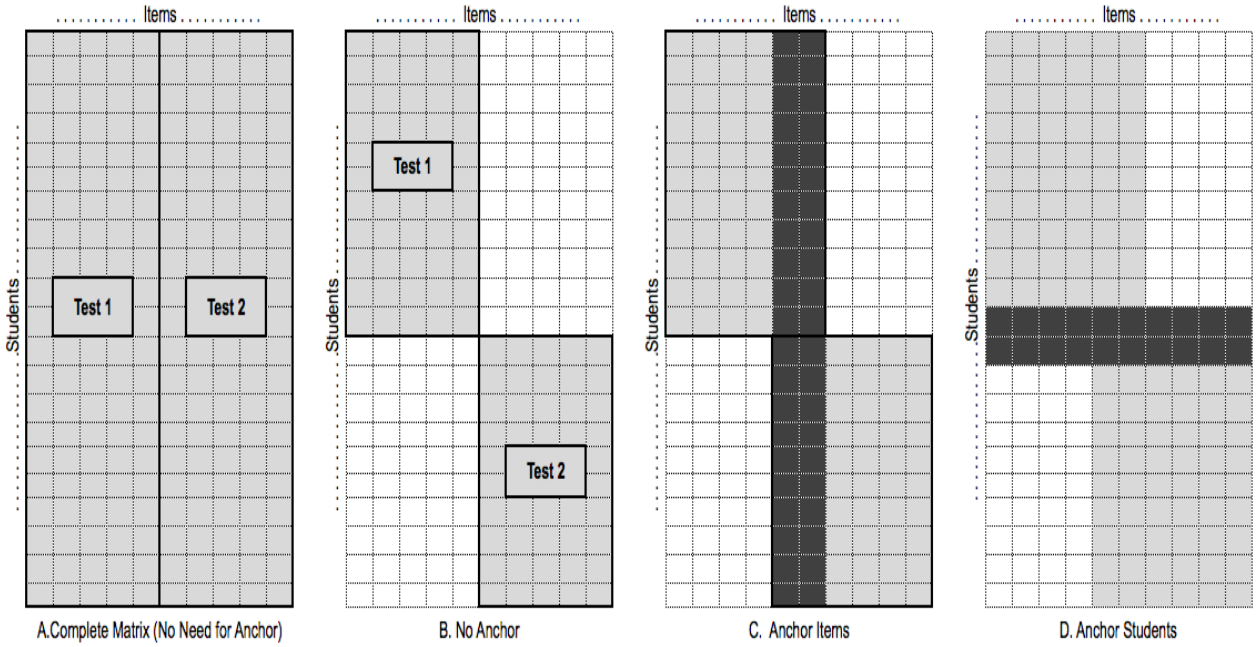

Figure 10. Item Response Data Matrices With and Without Overlaps

Figure 10C shows a popular anchoring design with common items, which are indicated with dark shade. In this design, a subset of items is included in both tests and is administered to all students. The link is established through students' performance on the common items. Similarly, Figure 10D shows another anchor design, which uses common people as anchor. In this design, a subset of students, shaded again in dark gray, take both test 1 and test 2. Again, the link between the two light gray areas of the matrix is established through the performance of the common students.

Now, let's consider a more complex situation, which is closer to the current study. In Figure 11 (see below), each row is a student, as in the previous figure, but now each column is a testlet with five associated items. Figure 11A shows that a response matrix in which testlets A through J were administered to a unique pair of students. Like Figure 10B above, there is no anchor in Figure 11A, therefore no link is established among the light gray areas of the matrix. In the current study, if I sampled only the responses to the fifth randomly given testlets, the data matrix would look like Figure 11A.

In Figure 11B, three testlets, A, F, and J, are used as anchors and every student in the matrix took at least two of these anchor testlets. Notice even though Figure 11B is an incomplete data matrix with many cells with left blank (i.e., missing data), the anchor testlets create chaining links that connect different parts of the matrix where the response data exist (i.e., the light gray

areas). This enables the placement of all testlets at different levels of difficulty onto the common scale. The current study used an anchor design similar to Figure 11B.

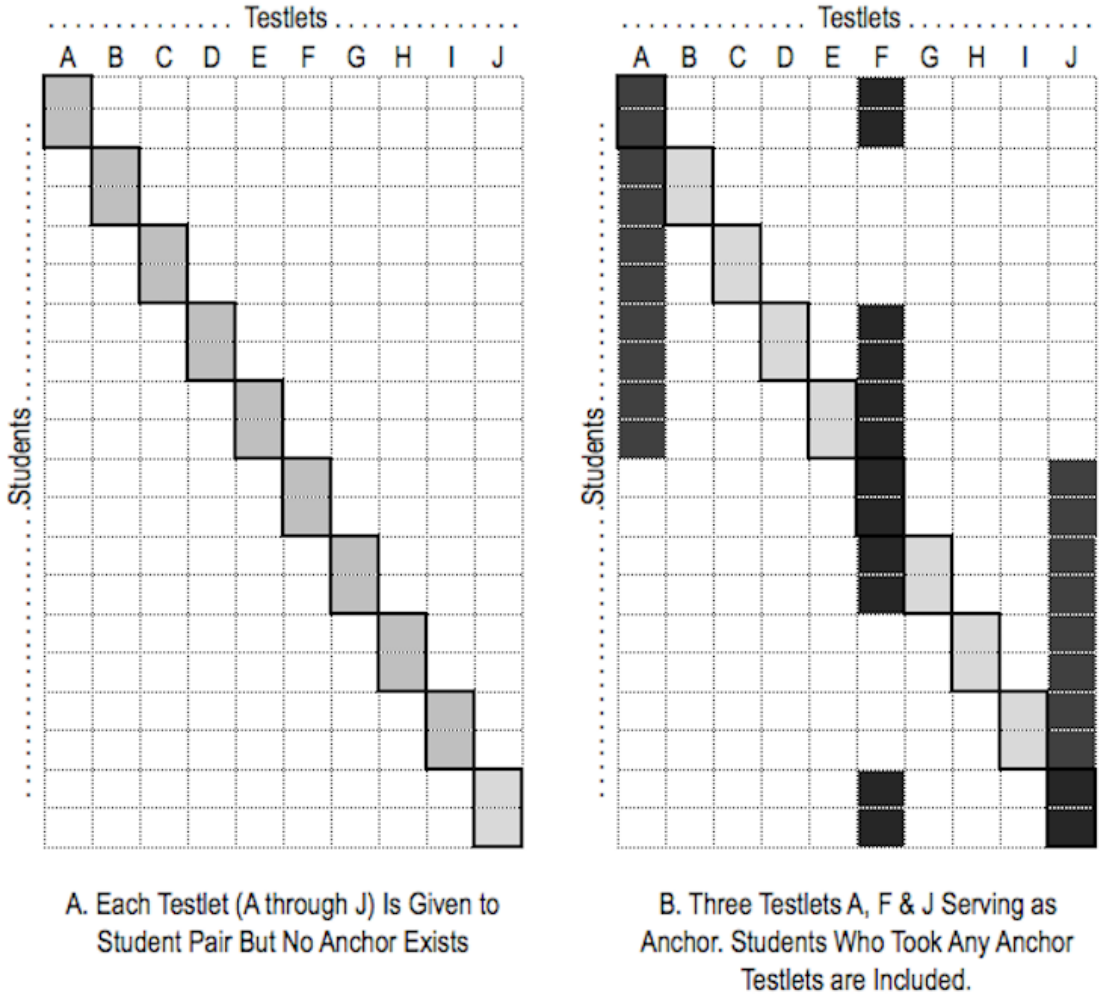

Figure 11. Item Response Data Matrices in a More Complex Case With and Without Overlaps

Specifically, for the current study, the original data set included a multi-state sample of 41,555 students from the United States and Canada, who took the ReadingPlus InSight assessment in the Winter of 2015-16. To develop a common vertical scale, seven testlets that provided sufficient overlap, dubbed "anchor testlets", were identified. These anchor testlets offered a total of 35 possible "common items", which amounted to about 15% of the total items examined. These anchor testlets are shown in Table 1 with gray highlights. As can be seen in the table, the anchor testlets were purposefully scattered across the 12 testlet levels determined by the test developer in order to cover lower-, middle-, and higher-levels of the testlets and to provide chaining linkages across the data matrix.

The final analytic sample included 10,547 students (out of the original 41,555) who took at least one anchor testlet as the second, third, forth, or fifth testlet within the five-testlet test administration. Students' responses to the anchor testlet(s) as well as to the fifth testlet were retained in the final dataset. Note that the fifth testlet could be any one of the 48 testlets in the

testlet bank, including the anchor testlets.[7] The students' responses to the first testlet were excluded as it was regarded a practice test; I assumed that in working on the first testlet, students familiarized themselves with the assessment, especially with the fact that they could not access the source passage while answering questions.

Table 1 shows the number of students who took each of the 48 testlets, tallied by the order in which the testlet was given. Note that for the anchor testlets (shaded in gray), non-zero numbers appear in all of the testlet order columns. In contrast, for the non-anchor testlets, only the fifth testlet column is filled with non-zeros. Typically, the analytic sample included students' responses to two testlets, but for some students, it included their responses up to four testlets (the latter group of students took multiple anchor testlets as their second through fifth testlets).

Table 1. List of 48 Testlets along with the Number of Students, by the Testlet Order

Testlet Title	Level	Testlet Order[a]				Total
		second	third	fourth	fifth	
Blobfish	1	0	0	0	208	208
Frog	1	0	0	0	235	235
Sand Dollar	1	0	0	0	247	247
Sea Star	1	109	117	157	259	642
Bread	2	0	0	0	244	244
Corn	2	0	0	0	229	229
Gum	2	0	0	0	269	269
Rice	2	0	0	0	260	260
Cave Art	3	0	0	0	225	225
Dance	3	0	0	0	219	219
Rain Sticks	3	0	0	0	240	240
Singing Bowls	3	0	0	0	224	224
Dragon Boats	4	0	0	0	204	204
Kite Fighting	4	0	0	0	228	228
Stick Juggling	4	0	0	0	253	253
Tribal Masks	4	0	0	0	219	219
History of Sports	5	547	528	564	198	1,837
Injury in Sports	5	0	0	0	238	238
Science in Sports	5	0	0	0	211	211
Technology in Sports	5	0	0	0	219	219
Black Holes	6	0	0	0	217	217
Comets	6	0	0	0	215	215
Solar Flares	6	485	546	560	213	1,804
Space Junk	6	0	0	0	228	228
High-Speed Boats	7	0	0	0	216	216
Mars Rovers	7	487	499	475	174	1,635
Self-Driving Cars	7	0	0	0	199	199
Unmanned Planes	7	355	339	330	203	1,227
Dream Catchers	8	0	0	0	203	203
Fish Rubbing	8	0	0	0	226	226
Paper Cutting	8	0	0	0	267	267

[7] However, if an anchor testlet was taken as the fifth testlet, another anchor testlet had to be taken as the second, third, or forth testlet.

Table 1 (continued)

Testlet Title	Level	Testlet Order				Total
		Second	Third	Fourth	Fifth	
Shadow Puppets	8	0	0	0	211	211
Fracking	9	0	0	0	251	251
Red Tide	9	0	0	0	213	213
Sailing Stones	9	0	0	0	161	161
Wildlife Crossings	9	0	0	0	229	229
Busby Berkeley	10	0	0	0	199	199
Duke Ellington	10	1,036	852	772	180	2,840
John Williams	10	0	0	0	211	211
Martha Graham	10	0	0	0	196	196
Bonsai	11	0	0	0	208	208
Dadaism	11	0	0	0	200	200
Kabuki	11	1,239	1,301	1,150	136	3,826
Mehndi	11	0	0	0	196	196
Coral Reefs	12	0	0	0	218	218
Deserts	12	0	0	0	252	252
Grasslands	12	0	0	0	263	263
Tundras	12	0	0	0	233	233
	Total	4,258	4,182	4,008	10,547	22,995

Note. Shaded in gray are the seven anchor testlets.

a. Responses to the testlet given in the first position of the five-testlet test administration were excluded as the first testlet was regarded as a practice (getting to know the assessment, especially the fact that the text was not available while answering questions).

Recall that the 2nd, 3rd, and 4th testlets were given according to the adaptive testing logic built into the assessment. To minimize the effect of this logic on the item difficulty estimates, I selected the testlets that showed the least discrepancy in item difficulty estimates between the calibration from the 2nd, 3rd, and 4th testlets with the adaptive logic, on the one hand, and the calibration just with the 5th, randomly selected testlets, on the other.[8]

Broadly speaking, there are two ways to calibrate item parameters for vertical scaling: (a) a single concurrent estimation using response data for all levels and (b) separate estimations for each test forms, followed by post-hoc linking of the results (e.g., linear transformation). The current study used the former as it intends to extract a single construct, in this case reading comprehension, across different levels, utilizing all of the available information in the data matrix. Additionally, the concurrent calibration is procedurally simpler and tends to produce accurate and stable results so long as the correct model is specified (Kolen & Brennan, 2004; R. J. Patz & Hanson, 2002).

Splitting the Student Sample for Cross Validation

The overall sample of 10,547 students was randomly divided into two samples of approximately equal size: sample 1 (n = 5,274) and sample 2 (n=5,273). Sample 1 was used as a

[8] To identify the seven testlets, two sets of item responses were calibrated concurrently with the Rasch model using the marginal maximum likelihood estimation. The first set was for the 240 items using the responses to the testlets given according to the adaptive logic while the second set was for the same 240 items but using only the responses to the fifth testlets (i.e., the randomly selected set). Selected as anchors were the testlets with the least discrepancy in item difficulty estimates between the two sets.

calibration sample for all statistical analyses while Sample 2 was used for subsequent cross-validation analyses, examining whether the results from the statistical analyses with Sample 1 would replicate with Sample 2.

Table 2 shows descriptive statistics for students' general vocabulary knowledge as measured by the vocabulary portion of the InSight assessment[9], grade levels, and the number of items responded, by two student samples. As expected, no statistically significant difference was found between the two student samples. Figure 12 shows the distributions of student grade levels, which are almost identical between the two samples. The total number of students per text ranged from 84 to 1,861 with a mean of 239.4 in sample 1 and from 77 to 1,964 with a mean of 239.6 in sample 2.

Table 2. Comparison of the Two Student Samples

	Sample 1 (n=5,274)		Sample 2 (n=5,273)			
	M	SD	M	SD	*t*	*p*
General Vocabulary Knowledge	5.42	1.58	5.45	1.58	1.20	0.23
Grade Level	6.78	2.38	6.82	2.42	0.82	0.41
Number of Items Responded	10.89	2.06	10.90	2.05	0.29	0.77

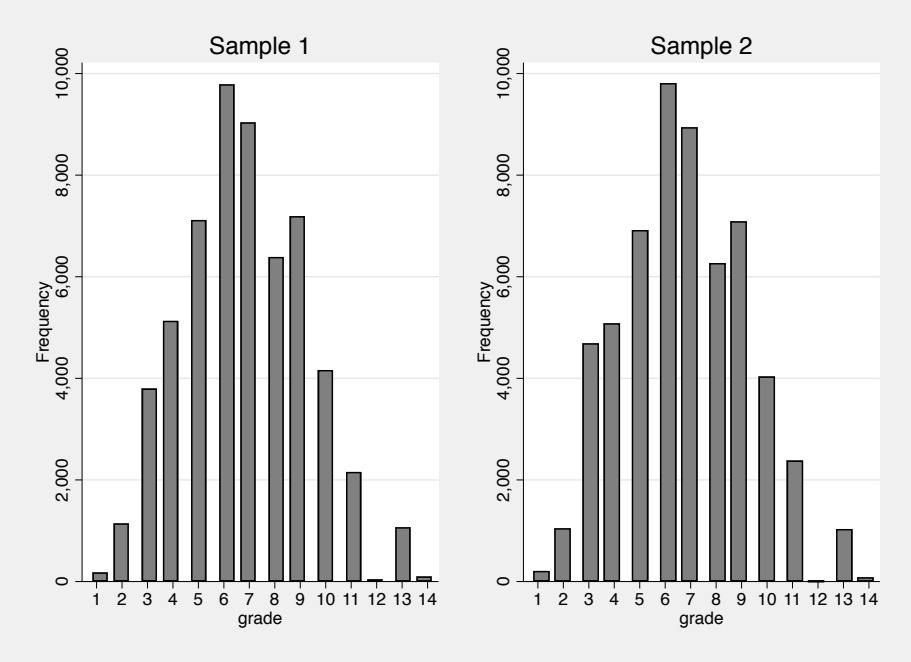

Figure 12. Distribution of students' grade level by two student samples.

Psychometric Models

All models used in this study are examples of Latent Regression-Linear Logistic Test Model (LR-LLTM). LR-LLTM are "doubly explanatory" because it has both item and personcovariates

[9] The vocabulary tested was based on graded list of words, which were not related to the passages given in the comprehension section of the assessment.

that explain item responses (De Boeck & Wilson (2004). As reviewed in Chapter 2, this model can be expressed as:

$$logit\left[\Pr\left(Y_{ip}=1\,|\theta_p,\beta_i\right)\right]=\sum_{j=0}^{J}\vartheta_j Z_{pj}+\epsilon_p-\sum_{k=0}^{K}\beta_k X_{ki}, \quad (8)^{10}$$

which can be easily understood as an extension of the Rasch model (equation 3 below) by the person side (θ_p) and the item side (δ_i) with equations 4 and 6:

$$logit(\Pr\left(Y_{ip}=1\,|\theta_p\right)=\theta_p-\delta_i, \quad (3)$$

$$\theta_p=\sum_{j=0}^{J}\vartheta_j Z_{pj}+\epsilon_p \quad \epsilon_p\sim N\left(0,\sigma_{\epsilon_p}^2\right), \quad (4)$$

$$\delta_i=\sum_{k=0}^{K}\beta_k X_{ki}. \quad (6)$$

The Rasch model (Equation 3) postulates that the log-odds of a person p correctly answering item i is a function of person p's ability (θ_p) and difficulty of item i (δ_i). Equation 6 extends the Rasch model on the item side by postulating that the item difficulty can be explained by a linear combination, over K realms, of item i's score for item feature k (X_{ki}) and its regression weight (or estimated fixed effect of the item feature, β_k). As noted before, Equations 3 and 6 together constitute the linear logistic test model (LLTM, Fischer 1973), which not only measures individual differences in the target person ability but also seeks to explain the item difficulty with item feature predictors that are assumed to underlie student's cognitive performance on a particular item. Note that in this study, the "item" features (X_{ki}) include both the passage features (e.g., mean sentence length in Lexile) as well as the task features (e.g., the number of falsifiable distractors).

As can be seen in Equations 8 and 4, LR-LLTM also extends the Rasch model on the person side with Z_{pj}being person p's value for person characteristics j (e.g., student's vocabulary level) , and ϑ_j being the regression weight of person characteristic j. Equations 1 and 4 together constitute the latent regression (LR) model (Van den Noortgate & Paek, 2004; Zwinderman, 1991). In LR-LLTM, Equations 1, 2, and 4 are run concurrently, enabling both measurement and explanation of individual differences along with item and person explanatory variables in one and the same analysis.

The equations (4)' and (6)' below shows how the Rasch model is doubly extended on the person and item side by including Lexile two factors, namely mean sentence length (MSL) and mean log word frequency (MLWF), as "item" features, and student general vocabulary knowledge (Vocab) as the person covariate:

$$\theta_p=\vartheta_1[\text{Vocab}_\text{p}]+\epsilon_p \quad \epsilon_p\sim N\left(0,\sigma_{\epsilon_p}^2\right) \quad (4)'$$

$$\delta_i=\beta_0+\beta_1[\text{MSL}_\text{i}]+\beta_2[\text{MLWF}_\text{i}] \quad (6)'$$

To investigate the modification of a particular passage's effect by the task features and the reader characteristic (Research Question 2), an interaction term was added for the item explanatory part of the model, as shown in Equation 6" below:

[10] Equation numbering continues from Chapter 2.

$$\delta_i = \beta_0 + \beta_1[\text{MSL}_i] + \beta_2[\text{MLWF}_i] + \beta_3[\text{MLWF}_i \times \text{Vocab}_p] \quad (6)''$$

In this particular example, the interact model examines whether the effect of MLWF was moderated by student general vocabulary knowledge. In this model, the base Rasch model part (Equation 3) and the person explanatory part (Equation 4') stay the same, and are concurrently run with Equation 6''.

Model comparisons. Model comparisons were conducted using Pseudo-R^2 proposed by Embretson (1983), which she calls the "delta fit index":

$$\Delta^2 = \frac{\ln L_0 - \ln L_m}{\ln L_0 - \ln L_S}, \quad (9)$$

where $\ln L_0$ is the log-likelihood for the null model, in this case a model with just an intercept (β_0). This can be interpret as a constant difficulty value for all items. In other words, the null model postulates that there is no difference among the 240 items in term of item difficulty. $\ln L_m$ is the log-likelihood for the model to be evaluated, which is the LR-LLTM specified in the previous section. $\ln L_S$ is the log-likelihood for the saturated model, that is, the Rasch-latent regression model which uses item dummies to represent all items as well as the student general vocabulary knowledge as the person covariate. Essentially, the denominator in Equation 7 shows the maximum amount of item difficulty that can be modeled by the latent regression which estimates difficulty for all items, while the numerator shows how much improvement the item feature predictors make compared to the null model that assumes equal difficulty for all items. The resulting ratio places the model to be evaluated on a scale from 0 to 1— comparable in magnitude to R^2 in ordinary least squares regressions; the value closer to 1 indicates better in terms of the explanatory power.

Additionally, standard model fit indices, Akaike's (1974) information criterion (AIC) and Schwarz's Bayesian information criterion (BIC, Schwarz, 1978), were examined. A model with a small value of AIC or BIC is preferred as having a better fit with the data. The two indices do not always select the same model as best fitting model. BIC is known to select simpler models with fewer predictors than AIC (Lin & Dayton, 1997). Further, for selected hierarchically nested models, the log-likelihood ratio (LR) test was conducted to compare a reduced model with a general model (the latter has more parameters). Statistical significance by the LR test means that the reduced model is rejected, favoring the general model.

Cognitive Variables

Passage features affecting the text representation phase. This study used several sets of passage features as "item" features, affecting the Text Representation phase of Embretson and Wetzel's processing model. This means that five items that were associated with the same source passage received a same value on a particular passage feature variable. For example, Lexile Text Analyzer gave the following two values to five items associated with "Sand Dollar", a first-grade passage: 8.4 as mean sentence length (MSL), and 3.72 for mean log word frequency (MLWF). Subsequently, both MSL and MLWF were included as the "item" features, representing Lexile model of text complexity. In addition to Lexile, three additional models of passage complexity were used in this study, namely (a) Gorin & Embretson's model consisting of three passage features, (b) Coh-Metrix eight text easiness components, and (c) Text-Evaluator's eight text complexity components. Table 3 provides a short description of each passage feature from these

three models (for details, see Gorin & Embretson, 2006; McNamara et al., 2014; and Sheehan et al., 2014; respectively).

Table 4 below provides descriptive statistics for a total of 21 passage feature predictors used in the study in their original unit/scale. In the actual analysis, these passage feature variables were standardized into z-scores except for the eight Coh-Metrix measures that were in the Z-score metric.

Table 3. Description of Text feature Variables in Three Text Complexity Models

Gorin & Embretson
Modifier propositional density is the number of modifier propositions divided by total number of words in text (Embretson & Wentzel, 1987). Gorin & Embretson (2006) used the number of adjectives divided by total number of words as proxy. The current study used Coh-Metrix occurrence score for adjectives (i.e., the number of adjectives per 1000 words) as proxy.
Predicate propositional density is the number of predicate propositions divided by total number of words in text (Embretson & Wentzel, 1987). Gorin & Embretson (2006) used the number of verbs divided by total number of words as proxy. The current study used Coh-Metrix occurrence score for verbs (i.e., the number of verbs per 1000 words) as proxy.
Text content vocabulary level is the average word frequency of the text. Embretson & Wentzel (1987) used Kucera-Francis (1967) index of word frequency. This study used mean log word frequency from Lexile
Coh-Metrix
Narrativity indicates the extent to which the text conveys a story, a procedure, or a sequence of events and actions with animate beings. Higher the narrativity, easier to comprehend.
Referential cohesion indicates the extent to which content words and ideas are connected with each other with use of noun phrases and other cohesion devices.
Syntactic simplicity is high when text includes fewer words and simpler and more familiar grammatical structures. In contrast, syntactically complex text requires the reader to hold many words and ideas in his/her working memory.
Word concreteness is high when words in the text are concrete, meaningful and evoke mental images. Conversely, it is low when the text includes more abstract words.
Causal cohesion indicates the extent to which clauses and sentences are linked with causal and intentional connectives.
Verb cohesion indicates the extent to which verbs overlap in a given text. More repeated verbs, more coherent a event structure that the text conveys.
Logical cohesion shows the extent to which explicit adversative/contrastive connectives (e.g., although), and additive connectives (e.g., moreover) are used to represent logical relations in the text.
Temporal cohesion indicates the extent to which temporal connectives (e.g., first, until) are present in a text, making it easy for the reader to develop situation model of the message conveyed in the text.

Table 3 (continued)

TextEvaluator
Academic vocabulary indicates the extent to which the language of a text exhibits characteristic of academic texts than of nonacademic texts such as fiction.
Word unfamiliarity indicates vocabulary difficulty based on several features such as word frequency and rare word measures.
Concreteness indicates the extent to which words in text evoke tangible images; measures combined into this component are all based on the MRC psycholinguistic database (Coltheart, 1981).
Syntactic complexity is composed of several features such as mean sentence length, average number of dependent clauses, average number of words before the main verb.
Degree of narrativity is comprised of three features: frequency of past perfect aspect verbs, frequency of past tense verbs, and frequency of third person singular pronouns.
Interactive conversational style indicates the extent to which a given text resembles spoken, conversational text, rather than to written, non-interactive texts.
Level of argumentation composed mainly of the frequency of concessive and adversative conjuncts, and the frequency of negations, indicating the amount of argumentation detected in a text.
Lexical cohesion composed mainly of repetition of content words across adjacent sentences and explicit connectives (e.g., consequently, for example).

Table 4. Summary Statistics for Text feature Variables (N=48 passages)

	Description	Scale/unit	Mean	SD	Min	Max
Gorin & Embretson Model						
1	Modifier propositional density	count of adjectives per 1000 words	91.58	20.49	43.10	136.36
2	Predicate propositional density	count of verbs per 1000 words	128.17	22.29	88.24	195.98
3	Text content vocabulary level	word frequency for content words	2.14	0.21	1.75	2.64
Lexile						
4	Ave. sentence length	count of sentences per passage	14.33	2.76	7.95	17.50
5	Mean log word frequency*	number of words per sentence	3.46	0.21	3.13	3.73
Coh-Metrix						
6	Narrativity*	z-score^{+}	-0.47	0.39	-1.12	0.13
7	Syntactic simplicity*	z-score^{+}	0.85	0.58	-0.25	1.76
8	Word concreteness*	z-score^{+}	1.61	0.57	0.53	2.44
9	Referential cohesion*	z-score^{+}	0.42	0.56	-0.78	1.26
10	Causal cohesion*	z-score^{+}	1.16	1.52	-1.37	4.05
11	Verb cohesion*	z-score^{+}	1.13	1.58	-1.17	3.58
12	Logical cohesion/ connectivity*	z-score^{+}	-2.32	1.2	-4.36	-0.08
13	Temporal cohesion	z-score^{+}	-0.08	0.86	-1.49	1.46
TextEvaluator						
14	Academic vocabulary	1 to 100 with 1 = least complex	34.10	18.40	19.00	71.00
15	Word unfamiliarity	1 to 100 with 1 = least complex	41.20	17.60	21.00	80.00
16	Concreteness*	1 to 100 with 1 = most complex	57.50	16.10	34.00	97.00
17	Syntactic complexity	1 to 100 with 1 = least complex	40.20	7.80	23.00	50.00
18	Degree of narrativity*	1 to 100 with 1 = most complex	39.60	20.40	4.00	66.00
19	Interactive conversational style*	1 to 100 with 1 = most complex	39.50	16.60	7.00	69.00
20	Level of argumentation	1 to 100 with 1 = least complex	40.60	21.00	7.00	74.00
21	Lexical cohesion*	1 to 100 with 1 = most complex	63.70	12.10	36.00	86.00

Note. Statistics are all in original unit. * are the variables that make text easier to process and comprehend.
$^{+}$ higher values indicate easier texts.

Task features affecting the response decision phase. In addition to the passage features, task features, which were found salient in the previous studies were used as item predictors. These task features were assumed to affect the Response Decision phase of the Embretson and Wentzel framework. Each of the 240 items was manually coded by human raters on three sets of the task features: *item-type/comprehension process*; *abstractness of information requested by the question*; and *falsification*. Figure 9 shows how these coding schemes were applied to a testlet called "Shadow Puppet".

The first coding system, *item-type/comprehension process*, was adapted from Ozuru et al., (2008), and is built on Kintch's construction and integration model of comprehension (Kintch 1988). This scheme classifies each of the 240 items into one of the four types in terms of cognitive processing required. The first type, *text-based (or literal recall) questions* ask for information that is explicitly stated within a single sentence in almost verbatim fashion, requiring minimal text processing. The second type, *restructuring/rewording* questions require students to identify the target information that cut across a few neighboring sentences within a paragraph. The information in the question is rephrased or restructured, thus require some amount of processing. The third type, *integration or bridging questions,* call for some degree of integration of information located across multiple paragraphs from the source passage. The last item type, *knowledge-based inference questions*, require information that is not explicitly stated in the source passage, thereby requiring students to bring in their prior knowledge to make inferences about the situation described in the passage.

The second coding scheme, *abstractness of information requested by the question*, was also adopted from Ozuru et al (2008) and was based on Monsenthal (1996). The underlying assumption of this scheme is that more extensive searching and more integration is required for questions that ask for abstract (e.g., a theme or lesson) rather than concrete information (e.g., a particular person). Thus the items requiring more abstract information for an answer is more difficult than those asking about more concrete things. Four levels exist in this scheme, ranging from highly concrete (e.g., specific animals, persons or things) to highly abstract (e.g., identification of equivalence/difference or a theme).

The third coding scheme examined the quality of distractors in terms of their falsifiability, following the prior studies that examined this construct (Embretson & Wentzel, 1989, Gorin & Embretson, 2006; and Ozuru et al., 2008). A distractor was falsifiable if the source passage provided explicit textual evidence against it (see the sample testlet and falsifiability codes for its answer choices in Figure 9). The number of falsifiable distractors was tallied per item, ranging from 0 to 3 (maximum possible was 4). Underlying assumption is that greater the number of falsifiable distractors is, the easier an item. Detailed descriptions of these coding schemes can be found in Appendix A.

To establish an inter-rater reliability, a second coder was trained on all the three coding schemes. I and the second rater each coded randomly selected 25% of items, resulting in above .80 inter-rater reliability on all the three variables. Additionally, the vocabulary demands of the distractors and of the correct answer choice in terms of the Flesch-Kincaid grade level were obtained using the Quantitative Discourse Analysis Package in *R* (Rinker, 2013). Descriptive statistics for these task level variables are shown in Table 5 and Table 6.

Table 5. Summary Statistics for Continuous Task Feature Variables (N=240 items)

	Mean	SD	Min	Max
Vocab level of question	4.99	3.09	-2.62	18.22
Vocab level of the correct answer	5.66	5.21	-3.40	26.49
Vocab level of distractors (mean)	5.29	3.13	-1.07	13.56
Number of falsifiable distractors	0.63	0.95	0.00	3.00

Table 6. Summary Statistics for Categorical Task Feature Variables (N=240 items)

Classification Scheme	Type of Questions	N	%
Item Type/ Comprehension Process	Text-based	18	9.09
	Reconstruct	67	33.84
	Integrate	61	30.81
	Knowledge-based	52	26.26
Abstractness	Highly concrete	45	18.75
	Somewhat concrete	60	25.00
	Somewhat abstract	84	35.00
	Highly abstract	51	21.25

Person covariate. Students' general vocabulary knowledge was used as a person covariate in this study.[11] Student's vocabulary was assessed within the ReadingPlus InSight assessment system, prior to the reading comprehension section. The multiple-choice vocabulary items asked students to select a word or phrase that most closely matches the meaning of the target word, which was taken from a list of 2,400 core academic words. Students' vocabulary ranged from 0 to 13 (in grade levels) with a mean of 5.4.

Analytic Process

Four broad categories of cognitive models of item difficulty were examined while controlling for students' general vocabulary knowledge, using the LR-LLTM. The first category comprised of the Text Representation (TR) models, each of which included a different set of text features that the literature suggests as affecting reading comprehension processes or readability of text. The second category of models included task feature predictors that are thought to affect the Response Decision (RD) processes in the context of multiple-choice RC tests. The third category combined the TR and RD models, examining the main effects of the passage and task features, while the fourth category of models added interaction terms to examine the modification of the passage's main effect, by the reader characteristic, the task characteristic (i.e., the item type), or the combination of both.

Initially, the saturated model (the latent regression) was implemented using ConQuest (Wu, Adams, & Wilson, 1998) to examine the Wright maps, especially for the distribution of item difficulty estimates. All the subsequent models were run using the *meglm* (multilevel mixed-effects generalized linear model) command in *Stata/SE 15.0*, which can handle continuous item predictor variables using the Marginal Maximum Likelihood (MML) as the estimation method.

[11] Ideally, other reader characteristics that affect comprehension (e.g., background knowledge and decoding skills) are used, but such data were not readily available for the current study.

Chapter 4. Results

This chapter presents results in five sections. First, results from the descriptive analyses are presented, including the pairwise correlations among the continuous variables. Second, the results from the models that examined the effects of passage features on item difficulty are presented. These models are referred to as the text-representation (TR) models as the predictors examined were hypothesized to affect the text representation phase of Embretson and Wentzel's model. Third, the findings from the models that analyzed the effects of the item/task features are presented. These models are referred to as the response decision (RD) models, reflecting the second stage of the Embretson and Wentzel model. Fourth, the findings from the TR and RD combined model are presented. The last section presents the results from the interaction models that examined the modification of passage feature effects by the reader and task features.

The sections 2 through 4 adress the first research question: Which set of text and task features best explain the variability in the difficulty of RC items, after controlling for student general vocabulary knowledge? The section 5 addresses the second research question: Are any of the text feature effects moderated by student general vocabulary knowledge and/or by task demands?

Descriptive Analysis

The Rasch model (without any person covariate) provided item difficulty estimates for 240 items. In this model, the mean student ability is constrained to 0 while the population SD ($\sqrt{var(\theta)}$) was estimated to be .68. The mean item difficulty was estimated to be .11 logits with a standard deviation of 0.66 and a range from -1.65 to 2.17. It is this variability in item difficulty that subsequent models sought to explain with various passage and task predictors while controlling for student's general vocabulary knowledge. Figure 13 shows the distributions of the ability estimates (on the left panel) and of the item difficulty (on the right panel; note that items on the x-axis are ordered by passage levels). Item difficulties are color coded to indicate the twelve passage levels as determined by the assessment developer. As can be seen in the figure, there is a general upward trend in the item difficulties as the passage levels increase along the x-axis, although there are substantial overlaps among the items from passages placed at neighboring levels. A rank-order correlation between the item difficulties and the passage levels was .76.

Table 7 provides an intercorrelation matrix between the item difficulty estimates and all of the continuous passage- and task-variables used in this study. As can be seen in the table, *academic vocabulary,* and *mean sentence length*, *word unfamiliarity*, and *syntactic complexity* have moderately positive correlations with the item difficulty (r=.64, .63, .60, and .56 respectively) while *mean* log *word frequency (MLWF)*, *narrativity*, and *verb cohesion* have moderately negative correlations (r= -.60, -.53, and -.51, respectively). In contrast, *argumentation*, *word concreteness*, *logical cohesion*, *causal cohesion,* and *falsifiability* have very low correlations with the item difficulty (r<.01, r=.05, .05, .03, and .03, respectively).

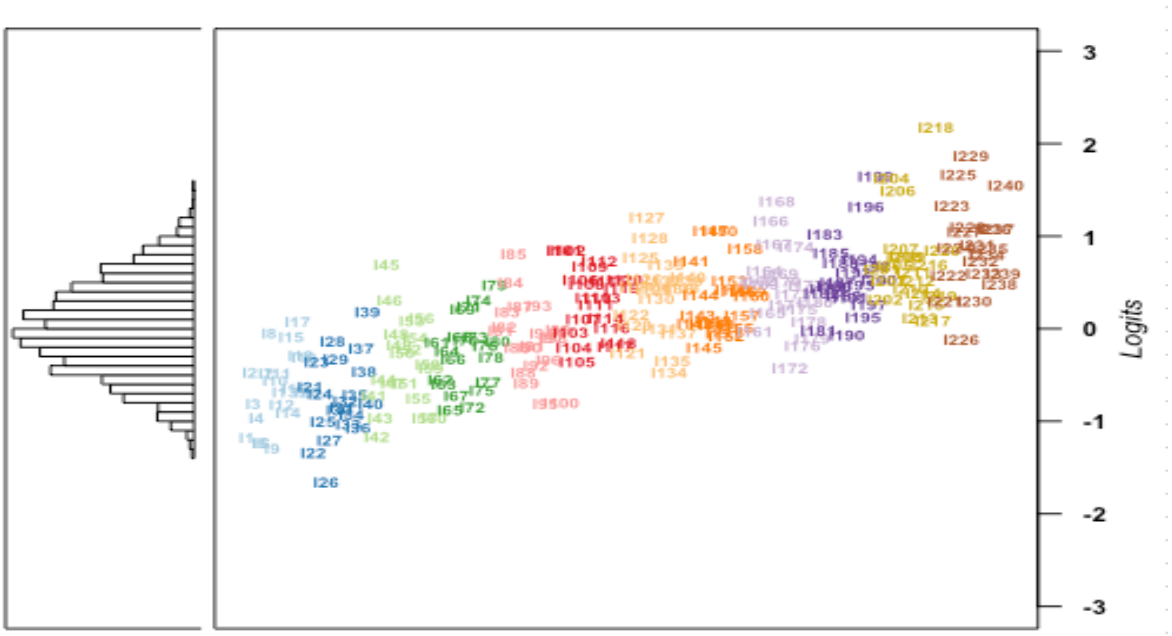

Figure 13. WrightMap from the Rasch model

Note. The figure shows the distributions of item difficulty for the 240 items (on the right panel) and of 5,274 students (on the left) from the Rasch model. The EAP (expected a posteriori) estimates were used for the ability. The items are color coded by the passage levels determined by the assessment developer.

Among the passage features, *mean sentence length* had a very high correlation with *syntactic complexity* (r=.94), which was expected as both are measuring similar constructs. *Mean sentence length* was also highly correlated with *word unfamiliarity* (r=.83), *mean word log frequency* (r =-.82), and *lexical cohesion* (r =.83). Similarly, *mean log word frequency*, *word familiarity* and *academic vocabulary* were all highly correlated from one another (r >.85 for all three possible pairs).

Among the task features, the vocabulary demand of the correct answer choice was correlated most highly with the item difficulty (r=.38). Naturally, this variable also had a moderate correlation with passage level vocabulary variables such as *academic vocabulary* (r=.52), *mean log word frequency* (r=.50), and *word unfamiliarity* (r=.50).

Table 7. Bivariate Intercorrelations among Item Difficulty and Passage and Task Characteristics

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	**Item Difficulty**	1.00																								
2	modifier prop dens[a]	.27	1.00																							
3	predicate prop den[a]	-.26	-.32	1.00																						
4	mean sent length[b]	.63	.32	-.24	1.00																					
5	MLWF[a,b]	-.60	-.45	.33	-.82	1.00																				
6	narrativity[c]	-.53	-.51	.46	-.64	.69	1.00																			
7	syntactic simplicity[c]	-.47	-.29	.39	-.71	.62	.48	1.00																		
8	word concrete[c]	.05	-.18	-.02	.15	-.12	-.19	-.08	1.00																	
9	referential cohesion[c]	-.28	-.22	.05	-.47	.51	.35	.29	.11	1.00																
10	causal cohesion[c]	.03	-.01	.05	.15	.02	-.02	.14	.26	.16	1.00															
11	verb cohesion[c]	-.51	-.11	.01	-.68	.72	.28	.39	-.25	.30	-.03	1.00														
12	logical cohesion[c]	.05	-.20	-.19	-.22	.19	.08	-.18	-.09	.37	-.40	.09	1.00													
13	temporality[c]	-.15	.08	.08	-.15	.02	-.01	-.18	-.09	-.01	-.14	.27	-.11	1.00												
14	academic vocab[d]	.64	.40	-.33	.84	-.87	-.72	-.73	.03	-.53	-.13	-.67	-.01	-.12	1.00											
15	argumentation[d]	.00	-.09	.07	.02	-.11	.12	-.03	.09	.12	.34	-.19	-.20	.07	-.11	1.00										
16	lexical cohesion[d]	-.15	-.07	-.13	-.31	.30	.10	.12	.24	.75	.16	.17	.29	-.08	-.27	-.04	1.00									
17	concreteness[d]	-.40	-.13	-.16	-.65	.48	.35	.42	.28	.33	-.15	.49	.22	.10	-.62	-.06	.37	1.00								
18	Interactive-conversa[d]	-.38	-.20	.20	-.40	.49	.45	.41	-.23	.05	.01	.38	.08	-.11	-.44	-.14	.01	.31	1.00							
19	narrativity[d]	.31	.18	.16	.40	-.35	-.23	-.18	.07	-.31	.12	-.35	-.19	-.20	.31	.22	-.24	-.36	-.44	1.00						
20	syntactic complexity[d]	.56	.27	-.15	.94	-.71	-.51	-.64	.06	-.37	.20	-.61	-.19	-.24	.72	.06	-.29	-.63	-.30	.42	1.00					
21	word unfamiliarity[d]	.60	.29	-.25	.83	-.92	-.69	-.67	.12	-.55	-.13	-.72	-.11	-.03	.92	.07	-.36	-.57	-.43	.36	.71	1.00				
22	vocab question[e]	.23	.14	-.01	.29	-.28	-.19	-.20	-.02	-.26	.03	-.20	-.14	.02	.29	.03	-.18	-.23	-.14	.21	.26	.31	1.00			
23	vocab corr answer[e]	.38	.26	-.09	.49	-.50	-.34	-.37	.10	-.25	.01	-.44	-.08	-.04	.52	-.01	-.09	-.33	-.18	.24	.41	.50	.22	1.00		
24	vocab distractors[e]	.08	.10	-.07	.26	-.20	-.35	-.15	.37	.00	.16	-.19	-.22	-.05	.22	-.03	.10	-.22	-.33	.08	.14	.19	.02	.17	1.00	
25	falsifiability[f]	.03	.04	-.15	.00	-.01	-.07	-.03	-.06	.06	-.09	.03	.07	.01	.01	-.05	.08	.01	-.10	-.08	-.02	-.02	-.18	-.09	-.12	1.00

Note. N=240, coefficients larger than .24 are significantly different from .00 at p<.05 after Bonferroni adjustment.

[a] Gorin & Embreton, [b] Lexile, [c] Coh-Metrix, [d]TextEvaluator, [e]Flesch Kincaid Grade Level, [f]number of falsifiable distractors

Effects of Passage Features (Text Representation Models)

Table 8 shows the pseudo-R^2 and the goodness of fit for the null model (M0), the saturated model (MS), and five additional models that included passage feature predictors that are thought to affect the text representation (TR) phase of reading comprehension. The Gorin & Embreton model (M1) with three passage feature predictors, namely *modifier propositional density*, *predicate propositional density*, and *content vocabulary*, accounted for 46% of variability in the item difficulty that could be modeled. The next model—the Lexile model (M2)—accounted for approximately 50% of variance with just the two traditional variables, namely *mean sentence length* and *mean log word frequency*. The Coh-Metrix model (M3) and the TextEvaluator model (M4), each with a different set of eight passage predictors, explained 51.8% and 51.3% of variance, respectively.

To examine how much explanatory power the Coh-Metrix model (M3) had above and beyond the Lexile model (M2), two predictors from Lexile and Coh-Metrix eight predictors were combined in M5. Similarly, the Lexile and TextEvaluator models (M2 and M4) were combined in M6 to examine the additional variance explained by TextEvaluator. As can be seen in Table 8, the Lexile and Coh-Metrix combined model (M5) had the most explanatory power among the six TR models examined, accounting for 54.5% of the information that could be modeled. AIC and BIC also indicated that this model (M5) was the best fitting model among the TR models examined. Further, the likelihood ratio tests indicated that M5 fits the data better than the Lexile model (M2, $\chi^2(8)= 107.22$, $p<.001$) or the Coh-Metrix model (M3, $\chi^2(2)= 59.78$, $p<.001$).

The results from this model indicate that the Coh-Metrix eight passage easiness factors increase the explanatory power by five percent points over the Lexile two-factor model. On the logit scale, the five percent points translate into .20 (3.85 × .05; 3.85 is the full range of item difficulty from the Rasch model without any person or item feature predictors). Thus the effect size is .28 (.20 ÷ .68; .68 is the population SD of student ability estimates from the Rasch model). Similar results were found in the cross validation analysis (see Table B-1 in Appendix for details).

Table 8. Comparison of Text Representation Models

Model	Pseudo R^2	Log-Likelihood	No. of Item Par*	AIC	BIC
M0: Null (Latent regression, LR)	.000	-38491.31	0	76988.62	77015.49
MS: Saturated (LR-Rasch)	1.000	-37056.47	240	74596.93	76764.95
Text Representation (TR) Models					
M1: Gorin & Embretson	.460	-37831.19	3	75674.38	75728.13
M2: Lexile	.494	-37781.79	2	75573.57	75618.37
M3: Coh-Metrix	.518	-37748.78	8	75519.56	75618.10
M4: TextEvaluator	.513	-37755.46	8	75532.92	75631.47
M5: Lexile + Coh-Metrix	**.545**	**-37709.29**	**10**	**75444.59**	**75561.05**
M6: Lexile + TextEvaluator	.521	-37743.71	10	75511.42	75627.88

Note. Bolded is the best fitting model among the five TR models in the table.
* Number of item parameters estimated. Note there is an additional person parameter (student's vocabulary level) in each model.

Table 9 shows parameter estimates from three TR models, M2, M3 and M5. These estimates can be interpreted in the same manner as the standardized regression coefficients (i.e., the beta

weights). That is, they show the effects of the predictors on item difficulty, controlling for those of all other predictors in the model. Bolded values in the table indicate that the significant effects were also found in the cross validation analysis in the same direction (see Table B-2 in Appendix for details). As can be seen in the table, the two text predictors in Lexile (M2) had statistically significant effects on item difficulty in the expected directions—the longer the average sentence length of a passage, the more difficult it was to successfully answer an accompanying RC item. This is indicated by the positive coefficient for *mean sentence length* (β_{msl} = .28, SE = .02) with the effect size (hereafter ES) of 0.41.[12] In contrast, a passage with more familiar words made it easier to correctly answer an accompanying RC item as indicated by the negative coefficient (β_{mlwf} = -.17, SE = .03, ES = -.25). Similarly, six of the eight Coh-Metrix "easiness" factors (M3) had statistically significant effects, mostly in the expected, negative direction (i.e., higher the easiness factor value, the easier the item). Of these easiness factors in the Coh-Metrix model, the largest effect was found with *syntactic simplicity* (β_{Synt} = -.40, SE = .03, ES = -.59).

Table 9. Parameter Estimates for Text Representation (TR) Models

	M2 Lexile		M3 Coh-Metrix		M5 Lexile + Coh-Metrix	
	Est.	SE	Est.	SE	Est.	SE
Fixed Effects						
Item						
Mean sent length	**.28*** **	**.02**			**.21*** **	**.04**
Mean log word freq	**-.17*** **	**.02**			**-.20*** **	**.03**
Narrativity			**-.37*** **	**.03**	-.06	.04
Syntactic simplicity			**-.40*** **	**.03**	**-.12***	**.05**
Word concreteness			**-.09*** **	**.02**	>.01	.02
Referential cohesion			.03	.02	.02	.02
Deep cohesion			**.03*** **	**.01**	>-.01	.01
Verb cohesion			**-.15*** **	**.01**	>-.03	.02
Logical cohesion			-.02	.01	.03*	.01
Temporality			**-.11*** **	**.01**	**-.09*** **	**.02**
Person						
Vocabulary level	**.41*** **	**.02**	**.34*** **	**.02**	**.35*** **	**.02**
Intercept	**-.09*** **	**.01**	**-.17*** **	**.05**	**-.21*** **	**.05**
Random Effects						
reader variance $\sigma^2_{\epsilon_\theta}$	**.44****	**.02**	**.40*** **	**.02**	**.40*** **	**.02**

Note. Est. columns show fixed or random effects in logit, analogous to the standardized regression coefficient (beta weights). SE stands for standard error. Bolded are the significant effects replicated in the cross-validation analysis. * p<.05 ** p<.01 *** p<.001

Similarly, the Lexile and Coh-Metrix combined model (M5) found the following five out of the 10 passage-predictors significant: *mean sentence length, mean log word frequency, syntactic*

[12] Recall that the population SD of the student ability estimates from the initial Rasch model (with no person or item covariates) was of .68. Using this SD, the effect size of mean sentence length is .28/.68 = .41.

simplicity, connectivity, and *temporality*.[13] Of them, all but one (*logical cohesion*) were consistently found to have significant effects on item difficulty in the same direction with the cross validation samples (for details of the cross-validation analysis, see Table B-2 in the Appendix).

As noted in Chapter 3, Coh-Metrix *syntactic simplicity* measures a similar construct as *mean sentence length* in Lexile; both constructs target at syntactic complexity. But Coh-Metrix *syntactic simplicity* reflects not only sentence length but also other factors such as the number of modifiers per noun phrase, the number of words before the main verb of the main clause, and similarity in syntactic structure across sentences. This explains why *syntactic simplicity* had a significant explanatory power in the combined model (M5, β_{syt} = -.12, SE = .05, ES = -.18), even after controlling for *mean sentence length*. *Temporality* was the other Coh-Metrix easiness factor significantly affecting the item difficulty in both analytical samples. This variable indicates the extent to which a passage includes linguistic markers that denote temporal consistency, including inflections and tense morphemes (e.g., "-ed", "is", "has) and verb tense and aspect (e.g., "has completed", "is completing"). *Temporality* also reflects time frames of unfolding events as indicated by adverbial phrases such as *after* and *in a minute.* Research (e.g., Zwaan & Radvansky, 1998) has shown that these temporal markers help the reader build more coherent mental representations of situations described in the text.

Rather unexpected was that in the combined model, Coh-Metrix *logical cohesion* reversed its directionality of effect from negative (β_{logi} = -.02, SE = .01, ES = -.03 in the Coh-Metrix model, M3) to positive (β_{logi} = .03 to , SE = .01, ES = .04). This means that more connective markers (e.g., *although*, *whereas*, *moreover*) the text has, more difficult for students to answer an associated RC item correctly. The expectation was that more connective markers the text has, easier for the reader to construct the mental representation of the text because the connectives make logical relationships more explicit. However, this reversed pattern was not replicated in the cross validation analysis: in fact, the cross validation analysis found a significant negative (rather than positive) effect of *logical cohesion* on item difficulty in line with the expectation (see Table B-2 in the Appendix).

Effects of Item/Task Features (Response Decision Models)

Now we turn to the second category of models: the Response Decision (RD) models. The RD models were composed of task features as predictors of the item difficulty, along with the student's general vocabulary knowledge as a control variable. Table 10 shows the pseudo-R^2 and the goodness of fit for the five RD models examined (M7-M11). The first RD model (M7) included *vocabulary demand of the question, the correct answer choice, and distractors*, as measured by Flesch Kincaid's grade level readability. This model accounted for 14% of the information that could be modeled. The *item-type/comprehension process* model (M8), with three dummy variables representing the four levels of cognitive processes required by questions, accounted for 9% of variance in the item difficulty. The next two models, one *for abstractness of information requested by a question* (M9), and the other for *falsifiability of distractors* (M10), accounted for 5.6% and 3.2% of the variance respectively. When all the response decision predictors were combined (M11), the model accounted for 25.4% of the variance, which is less than half of the best TR model (M5 combining the Lexile and Coh-Metrix text features) was able to account for. This indicates that the text representation processes influence item difficulty substantially more than the decision processes.

[13] To examine whether the order of predictors would change results, another model was run with Coh-Metrix predictors entered first followed by Lexile predictors. The model yielded the same exact results as M5.

Table 10. Comparison of Response Decision Models

Model	Pseudo R^2	Log-Likelihood	No. of Item Par	AIC	BIC
Response Decision (RD) Models					
M7: Readability of question/choices	.141	-38288.29	3	76588.58	76642.33
M8: Item type/comprehension process	.088	-38364.84	3	76741.68	76795.43
M9: Abstractness of info asked	.056	-3841.89	3	76833.78	76887.54
M10: Falsifiability of distractors	.032	-38444.85	1	76897.70	76933.53
M11: All RD predictors (M6-M9)	**.254**	**-38126.64**	**10**	**76279.28**	**76395.75**
TR + RD Combined Model					
M12: TR + RD predictors (M5+M11 without Falsifiability)	**.588**	**-37647.56**	**20**	**75339.12**	**75536.21**

Note. Est. columns show fixed or random effects in logit, analogous to the standardized regression coefficient (beta weights). SE stands for standard error. Bolded is the best fitting model among the five TR models in the table. * Number of item parameters estimated. Note there is an additional person parameter (student's vocabulary level) in each model.

Table 11 shows parameter estimates from the three RD models (M7, M8, and M11). The model with the *readability of items and answer choices* (M7) indicates that all the three readability predictors had statistically significant, positive effects on item difficulty (i.e., higher the vocabulary demands, more difficult the item). Of them, the *readability of the correct answer choice* had the largest effect ($\beta_{vocab.corr.ans}$ =.17, SE = .01, ES = .25) while that of the question and of the distractors each had less than one third of the impact in terms of the effect size ($\beta_{vocab.q}$ = .04, SE = .01, ES = .06; $\beta_{vocab.distr}$ =.05, SE = .01, ES = .07).

The next model, *item-type / comprehension-process* (M8), revealed a somewhat unexpected pattern: after controlling for other variables in the model, the questions that required *restructuring information within a paragraph* was not significantly different from literal recall questions. Further, *integrate/bridging inference* questions that require the reader to connect information across paragraphs turned out to be significantly easier than the *literal recall* questions ($\beta_{integrate}$ = -.08, SE = .04, ES = -.12). In contrast, *knowledge-base* inference questions, which require world knowledge outside of the source text, were statistically more difficult than literal recall questions ($\beta_{knowledge}$ = .24, SE = .03, ES = .35), and this latter pattern was in line with the expectation. The same patterns were found in the cross validation analysis (see Table B-4 in the Appendix). The literature suggests that *literal recall* questions should be the least cognitively demanding while *reword/restructure* items, *integrate/bridging inference* questions, and *knowledge-base* questions be increasingly more demanding in terms of cognitive processes involved (e.g., Anderson, 1972; Embretson & Wetzel, 1987; Hua & Keenan, 2014; Ozuru et al., 2008; Pearson, Hansen, & Gordon, 1979). However, this pattern was not found in the analyses.

Table 11. Parameter Estimates for Response Decision (RD) Models

	M7 Vocab Demand of Item & Ans. Choices		M8 Item Type / Comprehension Process		M11 All RD Predictors	
	Est.	SE	Est.	SE	Est.	SE
Fixed Effects						
Item						
Vocab Demand, Question	**.04****	**.01**			<.01	.01
Vocab Demand, Correct Answer	**.17*****	**.01**			**.20*****	**.01**
Vocab Demand, Distractors	**.05*****	**.01**			**.02****	**.01**
Item type (ref = text-base)						
reword/restructure			-.04	.03	**-.19*****	**.05**
integrate			**-.08***	**.04**	**-.18*****	**.06**
knowledge-base			**.24*****	**.03**	**.09*****	**.05**
Abstractness of Info (ref=highly concrete)						
somewhat concrete					**.30*****	**.04**
somewhat abstract					**.17*****	**.04**
highly abstract					**.15****	**.04**
Falsifiability					-.02*	.02
Person						
Vocabulary level	**.32*****	**.02**	**.26*****	**.02**	**.32*****	**.02**
Intercept	**-.21*****	.01	-.14	.03	-.11	.04
Random Effects						
Reader variance $\sigma^2_{\epsilon_\theta}$	**.42*****	**.02**	**.42*****	**.02**	**.42*****	**.02**

Note. Bolded are the significant effects replicated in the cross-validation analysis.
* p<.05 ** p<.01 *** p<.001

Finally, the results from the model that included all the response-decision (RD) predictors (M11) show that all but one task-feature predictor had a statistically significant effect on item difficulty. The exception was the *vocabulary demand of the question*, which showed an insignificant, close-to-zero-coefficient, after controlling for all other variables in the model ($\beta_{vocab.q}$ <.01, SE = .01). In contrast, the *vocabulary demand of the* correct answer and of the distractors remain significant, with the former having a relatively large effect on item difficulty ($\beta_{vocab.corr.ans}$ =.20, SE = .01, ES = .29; $\beta_{vocab.distr}$ =.02, SE = .01, ES = .03). Additionally, the two sets of the ordered categorical predictors, *item type/processing* and *abstractness of information requested by the question*, revealed somewhat unexpected patterns: consistent with the results from the earlier model (M7), *reward/restructure* items as well as *integrate/bridging inference* items were easier than literal recall questions as indicated by their negative coefficients ($\beta_{\text{restructure}}$ =-.19, SE = -.05, ES = -.28; $\beta_{integrate}$ =-.18, SE = .06, ES = -.27). This was contrary to what the literature suggests.

As for *abstractness of information*, the three coefficients were all positive and significantly different from zero, indicating that the questions that ask for highly concrete items (the reference category) were the easiest. However, the results did not match with the expected order of the

difficulty: the most difficult, according to the results, were the questions that ask for *somewhat concrete* items, followed by those asking for *somewhat abstract*. The questions that ask for *highly abstract* items turned out to be the second easiest, after controlling for all other predictor variables in the model. Post-hoc analyses indicated that the three coefficients were significantly different from one another.

Lastly, *falsifiability* had a small yet significantly negative effect on item difficulty ($\beta_{falsifi}$ = -.02, SE = .02, ES = -.03). This is in line with the expectation because it indicates that items become easier with an increase in the number of distractors that could be falsified with the information explicitly stated in the passage. In the cross validation sample, the same directionality of the effect was found but the effect was very small (less than -.01) and was not statistically significant (for details, see Table B-4 in the Appendix).

Effects of Passage and Item/Task Features (TR + RD Combined Models)

The third category of model combined the final TR model with the 10 passage predictors (M5) and the final RD model with nine task predictors (M11).[14] The last row in Table 10 above shows the pseudo-R^2 and the goodness of fit for this TR and RD combined model (M12). The model achieved the pseudo-R^2 of .588 and was the best fitting model among all the models examined so far, according to the BIC, AIC, and log-likelihood (see Table 10). The comparison of this value for the best TR model (M5, pseudo-R^2 =.545) indicates that the set of task variables contributes to additional five percent points increase in the amount of variance explained. This translates into the effect size of .28. Further, the likelihood ratio tests indicated that this combined model fit the data better than the final TR model (M5, $\chi^2(9)= 11.39$, $p<.001$) or the final RD model (M10, $\chi^2(9)= 856.36$, $p<.001$).

Table 12 shows the parameter estimates from the TR and RD combined model (M12). The same set of the four text feature predictors from the Lexile and Coh-Metrix combined model (M5) proved to significantly affect the item difficulty in the expected directions, after controlling for all other variables in the model. These text features were: *mean sentence length* (β_{msl} = .22, SE = .04, ES = .32)*, log mean word frequency* (β_{lmwf} = -.18, SE = .03, ES = -.27)*, syntactic simplicity* (β_{syt} = -.11, SE = .05, ES = -.16)*,* and *temporality* (β_{temp} = -.08, SE = .02, ES = -.12). The cross validation analysis also found that these four text features, along with *narrativity* and *connectivity*, had significant effects on difficulty (for details, see Table B-5 in the Appendix).

[14] The two Lexile predictors and the eight Coh-Metrix predictors were all kept intact, even though some of the Coh-Metrix's predictors did not have significant effects on item difficulty. This decision was made so that the structure of these two text complexity models were retained. In contrast, *falsifiability of distractors*–one of the task features–was dropped in the TR and RD combined model as it did not have a stable significant effect on item difficulty across the two analytical samples examined.

Table 12. Parameter Estimates for TR and RD Combined Models

	M12 TR + RD combined	
	Est.	SE
Fixed Effects		
Text Representation (TR)		
Mean sent length	**.22*** **	.04
Log mean word freq	**-.18*** **	.03
Narrativity	.07	.05
Syntactic simplicity	**-.11***	.05
Word concreteness	.03	.02
Referential cohesion	>.01	.02
Deep cohesion	-.02	.01
Verb cohesion	>-.01	.02
Logical cohesion	.03	.01
Temporality	**-.08*** **	.02
Response Decision (RD)		
Vocab Demand, Question	-.02	.01
Vocab Demand, Correct Ans.	**.03*** **	.01
Vocab Demand, Distractors	**-.04****	.01
Item Type (ref = literal-recall)		
reword/restructure	-.15***	.04
integrate	-.12**	.05
knowledge-base	>.01	.04
Abstractness of Info (ref=highly concrete)		
somewhat concrete	**.22*** **	.04
somewhat abstract	**.20*** **	.03
highly abstract	**.14****	.04
Reader		
General vocabulary knowledge	**.35*** **	.02
Intercept	**-.10****	.07
Random Effects		
Reader variance $\sigma^2_{\epsilon_\theta}$	**.41*** **	.02

Note. Bolded are the significant effects replicated in the cross-validation analysis.
* p<.05 ** p<.01 *** p<.001

As for the decision variables, the *vocabulary demand of the correct answer* had a small yet significantly positive effect ($\beta_{vocab.corr.ans}$ =.03, SE = .01, ES = .04) while *the vocabulary demand of the distractors* had a small negative effect ($\beta_{vocab.distr}$ = -.04, SE = .01; ES = -.06). The three dummy variables for the four ordered levels of *abstractness of information requested by the question* had significantly positive effects ($\beta_{sw.conc}$ = .22, SE = .04; ES =.32; $\beta_{sw.abs}$ = .20, SE

= .03; ES =.29; $\beta_{hi.abs}$ = .14, SE = .04; ES =.21), after controlling for all other predictors.[15] These patterns were also replicated in the cross validation sample. Additionally, as was the case with the final RD-predictor model (M10), *rewording/reconstruct* items as well as *integrate/bridging inference* items remained significantly easier than *literal-recall* items ($\beta_{reconst}$ = -.15, SE = .04, ES = -.22 ; $\beta_{integrate}$ = -.12, SE = .05; ES = -.18). In contrast, no significant difference was found between *literal-recall* and *knowledge-based* items. However, these patterns were not replicated with the cross validation sample: in the cross-validation analysis, no statistically significant differences were found among the four item types in terms of their difficulty (See Table B-5 in Appendix).

Taken together, the findings from the model building and the cross validation analyses suggest that the *literal recall* questions in the ReadingPlus InSight assessment are not easier than other question types even though they are commonly considered as least cognitively demanding. This result might be due, partially, to the fact that students could not refer back to the source passage while answering questions in the ReadingPlus InSight assessment. Without the ability to look back the source passage, the task of recalling specific, localized information in the text might be more challenging than or as challenging as, the other question types. This point will be further elaborated in the Discussion chapter below.

Modification of Text Effects (the Interaction Models)

As has been reported earlier, the following four passage features had statistically significant effects on item difficulty in both the model building and cross validation analyses: *mean sentence length (MSL)*, *mean log word frequency (MLWF)*, *syntactic simplicity (Synt)* and *temporality (Temp)*. The last set of models examined whether these main effects of passage features were moderated by a) the reader characteristic (i.e., student general vocabulary knowledge), b) the question type (a task characteristic), and c) both the reader and the task characteristics. Table 13 summarizes the model fit indices for all the interaction models examined. It shows that compared to the main-effects only model (M12), all but one interaction models significantly increased explained variance in item difficulty as indicated by increase in the pseudo R^2 value. In particular, one of the text-task interaction models (M19, including *temporality-item type* interactions) and three of the four three-way interaction models (M20, M21, and M23) increased the explained variance by more than three percentage points. Of them, a three-way interaction model involving *temporality* (M23) was the best fitting model in terms of the pseudo R^2, log-likelihood, AIC and BIC. This model explained 62.9% of variance in item difficulty, which is the largest variance explained among the 23 explanatory models examined in this study. Further, the likelihood ratio tests indicated that M23 fits the data better than the main effects-only model (M12, $\chi^2(10)= 99.18.22$, $p<.001$).

[15] Consistent with M11 (with all RD variables), the three coefficients for the three categories of the abstractness of information predictor were significantly different from one another.

Table 13. Comparison of Interaction Models by Pseudo R2, AIC, and BIC

Model	Pseudo R^2	Δ^+ Pseudo R^2	Log-Likelihood	No. of Item Par+	AIC	BIC
TR + RD Combined Model						
M12: TR + RD combined	.588	--	-37647.56	20	75339.12	75536.21
Text-Reader Interaction Models						
M13: M12 + MSL×Rvoc	.597	.009	-37634.19	21	75314.37	7552.42
M14: M12 + MLWF ×Rvoc	.596	.008	-37635.94	21	75317.88	75523.93
M15: M12 + Synt×Rvoc	.591	.003	-37643.77	21	75333.55	75539.6
M16: M12 + Temp×Rvoc	.588	.000	-37647.52	21	75341.04	75547.09
Text-Task Interaction Models						
M17: M12 + MSL×IType	.594	.006	-37638.68	23	75327.35	75551.32
M18: M12 + MLWF×IType	.602	.014	-37626.95	23	75303.89	75527.86
M19: M12 + Synt×IType	.591	.002	-37644.03	23	75338.06	75562.02
M20: M12 + Temp×IType	.623	.035	-37597.85	23	75245.70	75469.67
Text-Reader-Task Interaction Models						
M21: M12 + MSL×IType×Rvoc	.619	.031	-37602.47	30	75268.95	75555.63
M22: M12 + MLWF×IType×Rvoc	.616	.028	-37606.79	30	75277.58	75564.26
M23: M12 + Synt×IType×Rvoc	.604	.016	-37625.24	30	75314.48	75601.16
M24: M12 + Temp×IType×Rvoc	**.629**	**.041**	**-37589.04**	30	**75242.09**	**75528.77**

Note. + Change in Pseudo R^2 from the main-effects only model (M12) * p<.05 ** p<.01 *** p<.001
Bolded are the values that are the most extreme for the model comparison purpose: the largest for pseudo R^2 and change in pseudo R^2 while the smallest for log-likelihood, AIC, and BIC.

In the sections below, results from the following three classes of interaction models are presented: (a) the text-reader interactions, (b) the text-task interactions, and (c) the text-reader-task interactions.

The text-reader interactions. To examine whether each of the main text effects was moderated by students' general vocabulary knowledge, four separate interaction models were run. In these models, all other predictors in the TR and RD combined model (the main-effects only model, M12) were included.

These analyses found all but one text-reader interactions to be statistically significant: *MSL* and reader's vocabulary knowledge ($\beta_{msl\times rvoc}$= .05, SE =.01; ES = .07); *MLWF* and reader's vocabulary knowledge ($\beta_{mlwf\times rvoc}$ = -.07 , SE = .01, ES = -.10); and *syntactic simplicity* and reader's vocabulary knowledge ($\beta_{synt\times rvoc}$ =-.07 , SE = .03, ES = -.07). The interaction between *temporality* and general vocabulary knowledge was not significant ($\beta_{temp\times rvob}$ = -.01, SE = .013, ES = -.01). The same pattern was replicated with the cross validation sample.

These interaction effects are best interpreted graphically as shown in Figure 14. To generate these graphs, the text feature variables were set at two contrasting levels where a high value was defined as one standard deviation above their respective means while a low value was defined as one standard deviation below their means. Specifically, in each panel in the figure, a gap between the two lines—one for the high value of the text characteristic and another for the low value of the text characteristic—depicts differences in the expected probability of correct response between the

two levels of the text feature, as a function of student general vocabulary knowledge (the latter is a standardized score and is plotted along the x-axis). Hinge-like shapes along the lines indicate the 95% confidence intervals for the success-rate estimates.

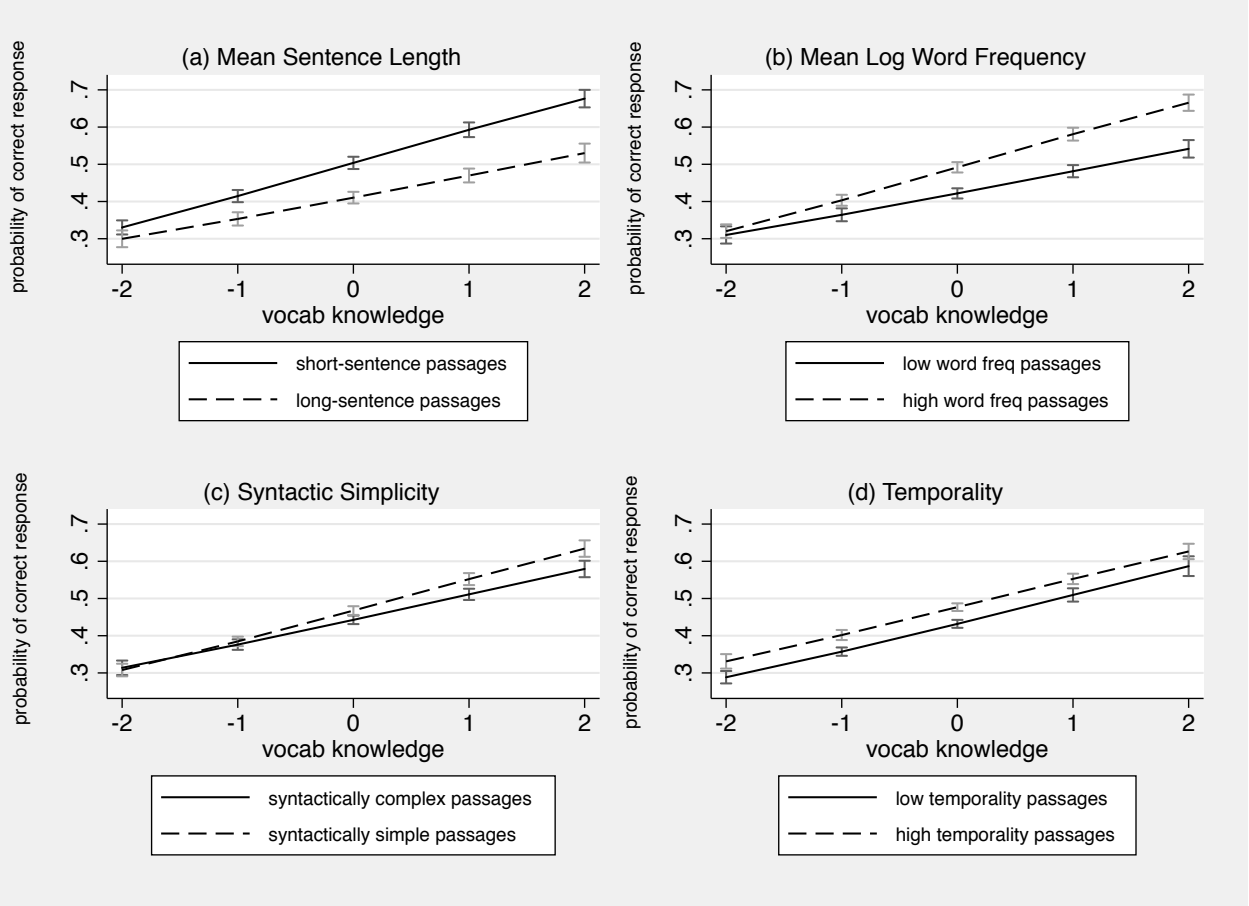

Figure 14. Four panels of line plots, each depicting interactions between general vocabulary knowledge and one of the four text features: (a) mean sentence length, (b) mean log word frequency, (c) syntactic simplicity, and (d) temporality, after controlling for all other text and item variables. All but panel (d) shows the modification of the text effects by general vocabulary knowledge, as evident with the widening of the gap between the two lines. Two levels of text feature variables were set at 1 SD above and 1 SD below their respective means.

As can be seen in the figure, all text features, except for *temporality,* provided greater help to students with higher level of general vocabulary knowledge than to those with a lower level, as evidenced by the widening of the gap between the two lines as student vocabulary level increases along the x-axis. Noticeably, MLWF (*mean log word frequency)* (panel b) and *syntactical complexity* (panel c) do not make a difference in the expected success rates for students whose general vocabulary knowledge score is -2, as indicated by the crossing of the two lines and the overlapping of the 95% confidence intervals. In contrast, the two lines for *temporality* (panel d) are parallel across the x-axis, indicating no differential impact of this text feature as a function of students' vocabulary level. Similar patterns were found in the cross-validation sample (see Figure B-1 in the Appendix).

The text-task interactions. We now turn to the text-task interactions, investigating whether the cognitive processes called upon by the four question types modify the main effect of the four text features, namely *mean sentence length (MSL)*, *mean log word frequency (MLWF)*, *syntactic simplicity* (*synt*) and *temporality (temp)*. The four question types investigated were: (a) text-base (or literal recall, the reference category), (b) reword/reconstruct, (c) integrate/bridging, and (d) knowledge-base, which the literature suggests are increasingly more demanding from (a) to (d). To facilitate the interpretation of the interaction effects, graphs were generated in a similar manner as in the previous section, with changes that the x-axis now represents the scores of text feature, rather than student vocabulary knowledge, and each panel has four lines representing each of the four item types. In these models, all other predictors in the TR and RD combined model (M12) were included.

All the four models, each examining the modification of one of the four text's main effects by the question type, found at least one statistically significant text-task interaction. Specifically, the significant interactions were found between: (a) *MSL* and all the three item-type dummy variables ($\beta_{msl\times reconst}$ =.17, SE = .04, ES = .25; $\beta_{msl\times bridging}$ = .14, SE = .05, ES = .21; $\beta_{msl\times knolwedge}$ =.18, SE = .05, ES = .26), (b) *MLWF* and all the three item-type dummy variables ($\beta_{mlwf\times reconst}$ = .17, SE = .04, ES = .25; $\beta_{mlwf\times integrate}$ = .18, SE = .04, ES = .26; $\beta_{mlwf\times knolwedge}$ =.23, SE = .04, ES = .34), (c) *syntactic simplicity* and the *knowledge-base* item type ($\beta_{synt\times knowledge}$ =.18, SE = .08, ES = .26), and (d) *temporality* and the two item types ($\beta_{temp\times integrate}$ =.21, SE = .04, ES = .31; $\beta_{temp\times knowledge}$ =.21, SE = .04, ES =.31).

As can be seen in Figure 15, the effects of *MLS* (panel a) and *MLWF* (panel b) had larger effects on *text-base* questions than other item types as indicated by their steeper slopes. Similarly, *temporality* (panel d) helped increase students' success with *knowledge-base* items as texts include more linguistic markers of temporality as indicated by the upward increasing line for the knowledge-base item type. In contrast, *temporality*'s effect appeared almost nonexistent with *reconstruct and integrate item types* as indicated by the relatively flat lines. With *literal-recall* items, temporality's effect was reversed, as indicated by a slightly downward sloping line. Finally, with *syntactic simplicity* (panel c), the line for *knowledge-base* items was noticeably flatter than the rest of the lines, indicating that this text feature did not have much effect on the items that rely on world knowledge outside of the source passage. Similar patterns were found in the cross validation analysis (for details, see Figure B-2 in the Appendix).

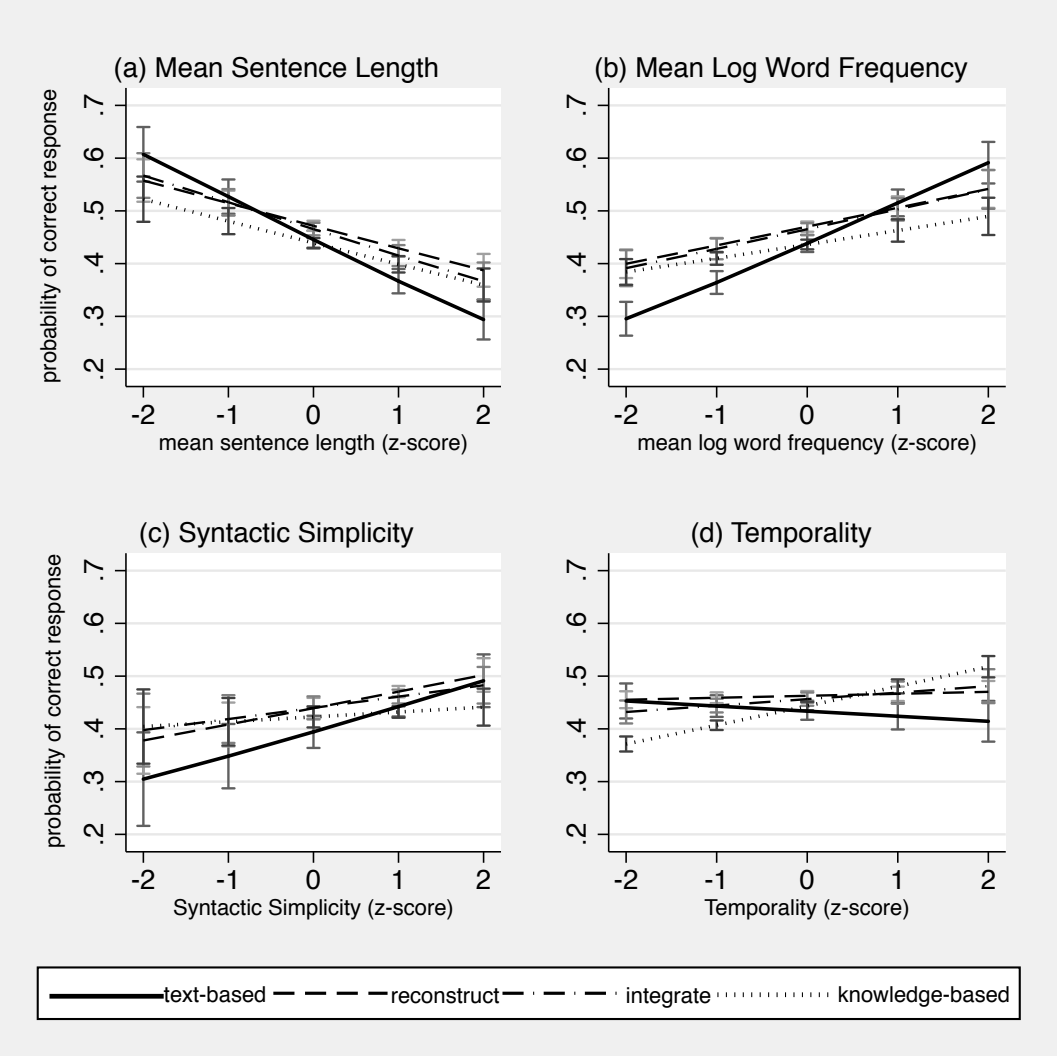

Figure 15. Four panels of line plots, each depicting interactions between the item type and one of the four text features: (a) mean sentence length, (b) mean log word frequency, (c) syntactic simplicity, and (d) temporality, after controlling for all other text and item variables. All but panel (c) shows the modification of the text feature by the item type.

The text-task-reader interactions. The text-task interactions reported in the previous section focused on the simultaneous text and task effects after controlling for students' general vocabulary knowledge as well as other text and task variables in the models. The last set of interaction analyses investigated whether the simultaneous text-task effects were moderated by students' general vocabulary knowledge. The models for these analyses included three-way interaction terms among the text, the task, and the reader as depicted in the RAND heuristic for reading comprehension. Four models were run, each involving three three-way interaction terms involving one of the four text features with main effects (i.e., *MSL, MLWF, syntactical complexity*, or *temporality*; these are all continuous variables), *student general vocabulary knowledge* (also a continuous variable), and one of the three *question types* (*reconstruct, integrate*, or *knowledge-base*, with the *text-base/literal recall* item type serving as a reference group, a categorical variable). All possible pairs of two-way interaction terms were also included in the models. Additionally, all other

predictors in the TR and RD combined model (i.e., the main-effects only model, M12) were included.

Of the four models examined, significant three-way interactions were found only in one model that included *temporality* as the text feature predictor. In that model, two of the three three-way interaction terms had statistically significant coefficients ($\beta_{temp\times rvoc\times reconst}$= -.12, SE = .05, ES = .18; $\beta_{temp\times rvoc\times integrate}$= -.12, SE = .05, ES = .18). The cross validation analysis found one additional three-way interaction significant (for details, see Figure B-3 in the Appendix).

Figure 16 shows the results from the three-way interaction models. In general, the text's effects for boosting in success rates were larger as students' vocabulary level increases along the x-axis across different question types. This is evident in the increasing gap between the two lines in each graph. However, one strikingly different display of the lines is observed for *temporality* (panel d) and *text-base question*, which require students to recall information in a single sentence in verbatim. And recall that this panel is based on the model that found the significant three-way interactions. It shows that passages with more temporal markers helped students with below-average general vocabulary knowledge answer the literal-recall questions. This is evident with the dashed line for the high temporality passages is above the solid line for the low temporality passages. However, for students whose vocabulary level is well above the sample average, it was actually the passages with fewer temporal markers that helped students' success with the text-base questions. Some researchers refer to a similar phenomenon as the "reverse cohesion effect" for readers with high background knowledge (O'Reilly & Mcnamara, 2007).

Another related finding is that high *temporality* texts with lots of time markers helped average and below-average vocabulary knowledge readers with *knowledge-base* questions, but this effect decreased as student vocabulary knowledge increased (notice that the gap between the two lines narrows as the general vocabulary knowledge increases along the x-axis). Interestingly, for students with a high vocabulary score of 2 (recall this is a z-score), *temporality* does not make much of a difference. This is evident in the graph with the point estimates for the success rates for the high and low temporality passages are close to each other and a large part of their 95% confidence intervals are overlapping.

Since the panel (d) of the Figure 16 showed almost completely overlapping lines for the *bridging/integrate* item type, this three-way interaction model was rerun by using *reconstruct* item type as the reference category (rather than the text-base item type). This analysis found only one of the three three-way interactions involving the text-base item type to be statistically significant ($\beta_{temp\times Rvoc\times textbase}$= -1.43, SE = .05, ES = 2.1). This finding was replicated in the cross validation sample. Taken together, these findings indicate that *temporality* had differential effects on the items that require students to recall information almost verbatim within a sentence, depending on students' general vocabulary knowledge: passages with more temporal linguistic markers helped low and mid vocabulary knowledge readers, but for high vocabulary readers, it was the low temporality (fewer time markers) passages that helped them more. Such reverse effects were not observed with any other item types examined in the study.

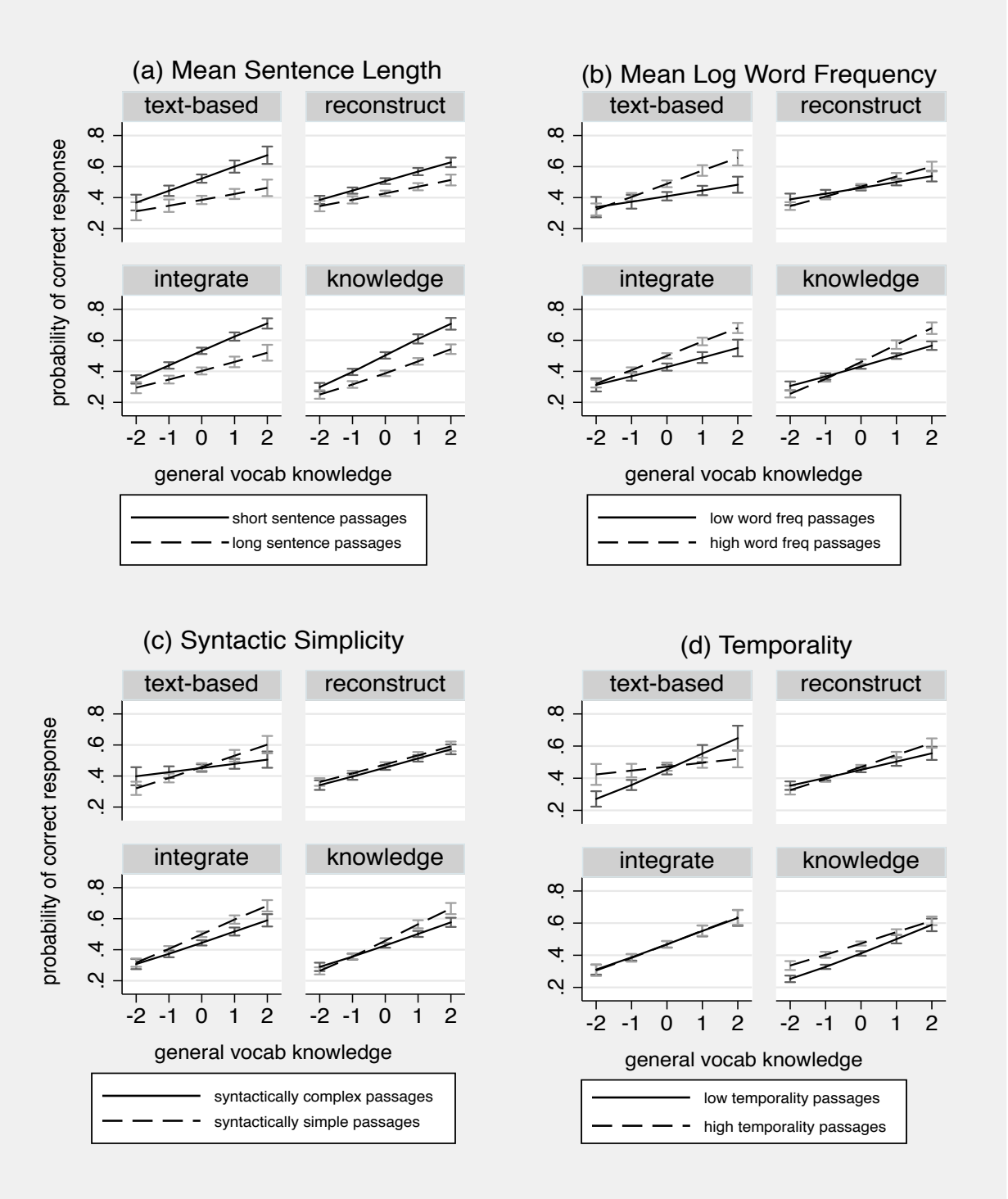

Figure 16. Four panels of line plots, each depicting three-way interactions among reader's general vocabulary knowledge, item type, and one of the four text features: (a) mean sentence length, (b) mean log word frequency, (c) syntactic simplicity, and (d) temporality, after controlling for all other text and item variables. Significant three-way interactions were found in the model involving temporality (panel d), which indicated the reverse effect of temporality for high-vocabulary knowledge readers on the text-base item type.

Chapter 5. Discussion and Conclusion

The main motivation for this study was to employ an advanced psychometric approach to directly model the features of the reader, the text, and the task in an effort to gain insights into their simultaneous effects on reading comprehension as postulated in the RAND "model" of reading comprehension. Specifically, I used explanatory item response models to investigate how the text and the task features influence difficulties of reading comprehension (RC) items, after controlling for students' general vocabulary knowledge. Additionally, I examined the text-reader, the text-task, and the text-reader-task interactions, to determine whether the effects of certain text features were moderated by the reader and the task characteristics.

While past studies have investigated the sources of cognitive complexity of RC assessments, these studies did not involve a wide range of readers and passages to sufficiently detect impacts of passage or task features. Such restriction of range in a predictor and an outcome variable is known to attenuate estimates of predictor weights, thus posing serious consequences to the conclusions that can be drawn. The current study overcame this hurdle by using unique item response data from a wide range of readers reading randomly selected passages, which as a set covered a wide range of levels from grades 1 through 12. Further, by vertically scaling the response data matrix on a common scale, the study sought to expand the current body of literature on how the reader, the text, and the task features may influence the reader's RC performance.

In what follows, three main findings are discussed: 1) passage features that affect students' success with the RC items, 2) item and task features that affect students' success, and 3) the interaction effects among the reader, the text, and the task characteristics.

Text & Task Features that Best Predict Item Difficulty

Among the individual text representation (TR) models with passage feature predictors examined, the Coh-Metrix model with eight text easiness factors explained the largest variance in the item difficulty (51.8%). The Lexile two-factor model also explained about 50% of the variance, and when these two models were combined, 54.5% of the total variance was explained. Of the 10 text feature variables included in the Lexile-Coh-Metrix combined model, four features, namely, *mean sentence length (MSL), mean log word frequency (MLWF), syntactic simplicity* and *temporality*, had significant unique effects on item difficulty in both the model building and the cross validation samples, while the effects of other correlated text features and the reader's general vocabulary knowledge were statistically controlled for. Interestingly, although Coh-Metrix *syntactic simplicity* and *temporality* show smaller effects than traditional two factors in Lexile (*MSL* and *MLWF*), they were consistently found to affect the difficulty of the RC items. This suggests that syntactic features that go beyond sheer sentence length (e.g., the number of modifiers per noun phrase, similarity in syntactic structure across sentences) do matter in students' success with the RC items. Further, it also indicates that not only the word and sentence-level features, but also the discourse level feature matters; in this case, what matters is the degree to which the passage included language markers that signify temporarily of events (e.g., temporal connectives such as *then, after, during;* temporal adverbs such as *in a moment*, *next day*, and consistency in verb tense and aspect such as *worked* vs. *is working*).

Among the response decision (RD) models with task feature predictors, the readability of the question and answer choices explained the largest variance in item difficulty (14%), followed by the item-type/comprehension processes (8.8%). The final RD model with all the significant item/task predictors explained 25.4% of the variance, which is only half of what the finale TR

model (with the Lexile and Coh-Metrix predictors) was able to account for. This finding suggests that in the ReadingPlus InSight assessment, the text processing variables (i.e., the passage features) affect the item difficulty more substantially than the response decision variables (i.e., the task /item features). This is contrary to the findings from the prior studies of other RC multiple-choice tests with older examinees. Embretson & Wetzel's 1987 study on the Armed Services Vocational Aptitude Battery (ASVAB) and Gorin & Embretson's 2006 study on the Graduate Record Examination (GRE) found that the difficulty of the RC items was primarily affected by the decision processes that involved mapping information between passages and answer options, rather than by the passage features that affect text processing.

This discrepancy in the findings can be explained at least in three ways. First, the ASVAB and the GRE might be focused more on verbal reasoning skills that involve the extensive mapping between the question, answer choices, and the source passage, than constructing a coherence representation of text meaning. Figure 17 shows sample passage and item from ASVAB and GRE. Indeed GRE's sample task appears to require more careful reasoning that involves going back and forth between the passage and answer choices than the ReadingPlus InSight tasks (for the latter, see Figure 9 in Chapter 3). The ReadingPlus InSight assessment, in contrast, is more focused on capturing students' ability to develop mental representations of the text rather than the verbal reasoning skills. This finding appears to fit the purpose of the ReadingPlus InSight assessment, which is to screen grades 1-12 students in need of individualized online reading intervention and place them into a particular reading level within the ReadingPlus' intervention program. In other words, arguably, the construct of RC measured by the ReadingPlus Insight may be different from those captured by the ASVAB or the GRE. Other researchers have made similar observations about the variability of RC constructs measured by different reading assessments, using the item difficulty modeling paradigm (e.g., Gorin & Embretson, 2006) as has been done in the current study, as well as correlating scores from multiple reading assessments (Keenan, Betjemann, & Olson, 2008).

Second, in the ReadingPlus Insight assessment, examinees could not access the source passage while answering questions. This is a stark difference from other, in fact the most, RC assessments examined in the previous studies of item difficulty modeling. In the earlier studies, the text was made available throughout the test administration. Because of this setup, the ReadingPlus Insight assessment requires students to rely on their memory of the text (or the representation of the text) when answering an RC question, more so than the other RC assessments that use the with-text administration. In fact, some researchers call this type of comprehension as memory-based comprehension (Artelt, Schiefele, & Schneider, 2001). In the memory-based comprehension assessment like the ReadingPlus Insight assessment, it is natural that students' success with answering RC items depends more on the passage features that affect the construction of mental models of text than the task/item features that affect the response decision process. Proponents of without-text administration, like the ReadingPlus Insight assessment, argue that the lack of text access enables one to capture automatized comprehension processes such as word decoding, knowledge activation, and inferencing to form a coherent model of text meaning, without making the reader engage in more effortful, non-linear text processing as well as in test-taking strategies (Artelt et al., 2001; Higgs, Magliano, Vidal-Abarca, Martínez, & McNamara, 2017). The findings from the current study suggest that the test administration format, specifically the availability of text while answering a question, may change the nature of a reading task, thereby changing the underlying RC construct being measured.

ASVAB	GRE
Nations are political and military units, but they are not necessarily the most important units in economic life, nor are they very much alike in any economic sense. All that nations really have in common is the political aspect of their sovereignty. Indeed, the failure of national governments to control economic forces suggests that nations are irrelevant to promoting economic success. Question: According to the paragraph, the economic power of nations is: A. controlled by political and military success B. the basis of their political success C. limited to a few powerful nations D. relatively unimportant	Reviving the practice of using elements of popular music in classical composition, an approach that had been in hibernation in the United States during the 1960s, composer Philip Glass (born 1937) embraced the ethos of popular music in his compositions. Glass based two symphonies on music by rock musicians David Bowie and Brian Eno, but the symphonies' sound is distinctively his. Popular elements do not appear out of place in Glass's classical music, which from its early days has shared certain harmonies and rhythms with rock music. Yet this use of popular elements has not made Glass a composer of popular music. His music is not a version of popular music packaged to attract classical listeners; it is high art for listeners steeped in rock rather than the classics. The passage addresses which of the following issues related to Glass's use of popular elements in his classical compositions? A. How it is regarded by listeners who prefer rock to the classics B. How it has affected the commercial success of Glass's music C. Whether it has contributed to a revival of interest among other composers in using popular elements in their compositions D. Whether it has had a detrimental effect on Glass's reputation as a composer of classical music E. Whether it has caused certain of Glass's works to be derivative in quality

Figure 17. Sample passages and items from ASVAB and GRE

Third, this study analyzed student responses to a wide range of reading passages designed for grades 1 through 12, which were intentionally designed to vary in vocabulary demand. This variability in the source passages may have helped for some of the text features to reach a greater explanatory power than the previous studies. In contrast, the prior studies analyzed student responses to an assessment that was targeted at a narrower range of examinees: ASVAB is for military applicants (Embretson & Wentzel, 1987); the GRE is for graduate school applicants (Gorin & Embretson, 2006); and the GMRT-RC has grade-band specific forms such as one for grades 7-9 and another for grades 10-12 (Ozuru et al, 2008; Kulesz et al., 2016). In fact, Gorin and Embretson (2006) admit that the GRE passages they analyzed did not vary very much in propositional density and content words. They suspect this might have contributed to the lower explanatory power of the text features' on item difficulty than the item features.

Interaction Effects

The current study uncovered several small but significant interaction effects among the reader, the task, and the text. Specifically, the interaction analyses targeted the four text features that consistently had significant main effects on the difficulty of RC items in both the model building and the cross validation samples. These text features were *mean sentence length*, *mean word frequency*, *syntactic simplicity*, and *temporality*. Of them, all but one (i.e., temporality) interacted with students' general vocabulary knowledge. In all three cases, a more positive level of the text feature (i.e., shorter sentences, more familiar words, or simpler syntactic construction) boosted high vocabulary knowledge students' performance at a greater extent than low vocabulary knowledge students, after controlling for all other variables in the model. These findings suggest that students with greater vocabulary knowledge benefit more from the positive affordances of these text features than students with limited vocabulary knowledge. This is hardly surprising given that a strong relationship has been found between reading comprehension and vocabulary knowledge in the literature (Carroll, 1993; Stahl & Fairbanks, 1986; Tannenbaum, Torgesen, & Wagner, 2006). However, such an interaction effect was not found between students' vocabulary knowledge and text's *temporality*— the extent to which a passage uses the consistent verb tense and aspect as well as temporal connectives such as *and then*, *after*, *during* that help the reader establish a mental sequence of events in the text. As will be described below, *temporality* behaved differently as compared to the other text features examined in the interaction analyses.

As for the text-task interactions, all but one of the four text effects, namely, *mean sentence length, mean log word frequency,* and *temporality*, were significantly moderated by item type. Specifically, the effects of *sentence length* and *word frequency* were larger with the *text-base* items than the other item types (i.e., reconstruct, integrate, and knowledge-base items). In contrast, the effect of *temporality* was most pronounced with *knowledge-base* items that require students to make inferences using their background knowledge, while it had no apparent positive apparent effect on other item types. These findings suggest that shorter sentences and familiar words helped the reader recall specific information localized within a single sentence in the text even when the text was not accessible at the time of question answering, while temporal linguistic markers made the text more considerate for the reader to activate their background knowledge.

Further, one of the three-way interaction models revealed the reverse effect of *temporality* for high vocabulary knowledge students: these students performed better on the literal-recall questions when the passages had fewer temporal markers while their peers with low vocabulary knowledge did better with passages with more temporal markers. Interestingly, this effect was confined only to the text-base item type; other items types did not show this pattern. Past experimental studies have found similar reverse effects of cohesion with middle school students (McNamara, Kintsch, Songer, & Kintsch, 1996) as well as college students (McNamara, 2001; McNamara & Kintsch, 1996). In these studies, students with higher background knowledge (rather than vocabulary knowledge) performed better on less cohesive passages as marked by low argument overlap and causal cohesion, while students with lower background knowledge benefited from highly cohesive passages. Researchers explained that when the text is too explicit about the relationships among ideas, it prevents high knowledge readers from making knowledge-based inferences to develop a coherent, and perhaps enduring, understanding of the text. In other words, highly-cohesive texts did not call for active processing of the text. In contrast, low-cohesion texts necessitated high knowledge readers to actively use their background knowledge to fill conceptual gaps in the texts, resulting in better integration of texts into their knowledge.

In a similar way, in the context of the current study, low-temporal cohesion may have encouraged high vocabulary knowledge students to more actively engage in the text processing, by making inferences using their background knowledge. This deeper processing, in turn, is likely to have helped them construct a more coherent situation model, helping them recall specific localized information in the text even when the text was not available during the question answering stage. It is not clear, however, why this reverse temporality effect was limited to the text-base items and did not extend to situation-model level questions such as ones that call for integrating information across paragraphs and drawing on background knowledge. The findings from the prior studies are mixed on this point: like the current study, McNamara (2001), Britton & Gülgöz (1991) and Magliano et al. (2005) reported that the reverse effect was found only with the text-base questions, while McNamara & Kintsch (1996) and McNamara et al (1996) found the reverse effects only with the situation model level questions.

Limitation of the Study

There are several limitations in this study that should be addressed in future research. First, the length of passages in the ReadingPlus InSight Assessment were rather short, which ranged from 168 to 282 words with the mean of 238.81 across 48 passages. Thus, the passages may not have sufficiently varied in their discourse-level features. Indeed, little variability was observed in the descriptive statistics for some of the cohesion features such as referential cohesion and lexical cohesion relative to other text features. Had the response data from longer assessment passages examined, results might have been different. On the other hand, the ReadingPlus' assessment passages were not remarkably shorter than the Gates-McGinite Reading Tests-Reading Comprehension (GMRT-RC) that had been repeatedly examined in the previous studies. For example, Kulesz et al. (2016) reported that GMRT-RC's passages for Grades 7-9 and Grades 10-12 were all shorter than 200 words.

Second, this study is inherently correlational in nature therefore causal inference about the effects of the passage, the item, and the reader characteristics on reading difficulty is limited. The current and most of the past studies have retrofitted the cognitive processing models (e.g., Embretson & Wentzel's processing model) to existing RC items as well as to their associated passages. As such, the passages and item features were not experimentally controlled. Statistical control is typically used to understand the unique influence of predictors on the difficulty of RC items by holding other variables in the model constant. However, it is unlikely that all potentially confounding variables were separated out through this process. Additionally, statistical control requires simplified assumptions in that effects of confounding variables are linear and additive. The current study included product terms among some of the predictors to explore the text-task-reader interactions. However not all complex relationships among the variables were likely to be accounted for. Future research should experimentally examine the contributions of the text and item features by systemically designing the texts and the items that vary on the features of interest. Such experimental research would allow explicit hypothesis testing (Gorin, 2005).

Third, general vocabulary knowledge was the only reader factor available for the study. While it may be a proxy for some of the key reader characteristics (e.g., background knowledge, fluency) previous studies have shown a host of other reader characteristics that account for student RC performance, such as word recognition and decoding (Cutting & Scarborough, 2006; Torgesen, 2000), oral language (Hoover & Gough, 1990), prior knowledge of passage topic (Miller & Keenan, 2009), working memory (Cain, Oakhill, & Bryant, 2004), executive function (Miller et al., 2014), higher-order skills such as inferencing, planning, and organizing (Rapp, Broek, McMaster,

Kendeou, & Espin, 2007), and affective attributes such as interest (Kirby, Ball, Geier, Parrila, & Wade-Woolley, 2011) and motivation (Guthrie et al., 2004). Ideally, future research collects more information about readers so that a wider array of reader factors could be investigated.

As for the psychometric modeling, the current study does not account for the nesting structure of items within passages. This means that possible dependencies among the items that share the common passage are ignored because the model assumes items are independently contributing to the difficulty estimates. The consequence of this is that estimates of regression weights and their standard errors are biased (the latter is typically underestimated while the former can be over- or under-estimated), while reliabilities are likely to be overestimated (Wang, Cheng, & Wilson, 2005).

A further technical limitation is that the models used in this study all assume item difficulties were perfectly predicted by a linear function of item and passage features. In other words, it was assumed that variance in item difficulties would be perfectly explained by the item-feature predictors. However this is an unrealistic assumption because the predictive linear function and its underlying substantive theory are never perfect and the item difficulties might be a random variable (De Boeck, 2008). The current study sought to relax this assumption by simultaneously allowing for residual variation in the item difficulties. In the psychometric literature, this extended model is called the LLTM with error (LLTM+e; De Boeck, 2008a; Janssen et al., 2004). Unfortunately, the LLTM+e did not converge, most likely due to a large amount of missing values in the data matrix (recall no students took all 48 testlets). The extant literature has pointed out that the LLTM+e is demanding in terms of estimation because both the persons and the items are treated as random, making it as a crossed random effects model. Future studies should follow a better data collection design that allows vertical scaling while minimizing the amount of missing values in the data matrix.

Implications for Future Research

There are several implications for future research in this area. First, future research should experimentally investigate the effects of text and task features on difficulty by manipulating a particular configuration of text or task features while holding other features constant. Such an experimental study will allow causal inference about the effects of manipulated text and item features on item difficulty. The current correlational study has offered an initial set of text and task variables to be considered as sources of RC difficulty. It will be interesting to examine whether the results from the current study will be replicated, especially the reverse effect of temporality for high vocabulary knowledge readers. It is imperative for such an experimental study to develop longer passages that would enable a more complete examination of the discourse-level features such as various cohesive measures provided by Coh-Metrix.

Second, future studiwa should use a sound design for data collection and linking so that a resulting response data matrix would involve much fewer instances of missing values. Such a design would likely allow the use of a more complete psychometric model like the LLTM plus the random error term, which relaxes the unrealistic assumption with the LLRM (i.e., item predictors perfectly predicts item difficulty). The data collection should also involve a wider range of reader characteristics that correspond to the passage and the task demands, such as background knowledge, working memory, syntactic knowledge, inference generation, and comprehension monitoring. Additionally, non-cognitive reader attributes such as interest and motivation would be informative.

Third, the current study has indicated that the latent construct of RC measured by the ReadingPlus Insight assessment might be different from other standardized RC assessments such as

the GRE and GMRT-RC— the very tests that prior studies investigated—primarily because the assessment does not allow students to look back the source passage when answering questions. It would be worthwhile to empirically investigate the effect of the source text's availability, examining whether the explanatory power of the task features, which are hypothesized to affect the response decision stage, increases when the passage is available during question answering. Further, the verbal protocol and/or the eye-tracking paradigm could be applied to examine whether the availability of text would change reading processes. Such a study would contribute to a small yet growing set of studies that have uncovered differences among RC construct as measured by different assessments, even with a slight change in the test design (Gorin & Embretson, 2006; Keenan et al., 2008; Rupp, Ferne, & Choi, 2006; Svetina, Gorin, & Tatsuoka, 2011).

Lastly and most important, future studies should develop a cognitive model of student performance on a wider range of RC tasks that call for more complex processing such as those included in the cognitive targets in the National Assessment of Educational Progress (NAEP, National Assessment Governing Board, 2015). Examples of the cognitively demanding tasks include (a) making complex inferences from multiple passages and (b) evaluating the author's craft and perspectives. Many of these tasks go beyond the multiple-choice item format that the Embretson and Wentzel's processing model can account for. Subsequently, the reader's thinking and problem-solving processes need to be specified for such cognitively demanding tasks, which would change what components and predictors need to be included, especially in the response-decision phase of Embretson and Wentzel's model. Ultimately, the cognitive model should facilitate the design of an assessment that would elicit the target reading behaviors and processes, which in turn would allow diagnostic inferences about readers' strengths and weakness. Additionally, the cognitive model should facilitate the description, explanation, and prediction of readers' performance in details, going beyond just a single number or label. Further, to more fully reflect the spirit of the RAND heuristic of reading comprehension, socio-cultural contextual factor(s) could be brought to bear in the analysis. In the case of the ReadingPlus Insight Assessment, time of the test administration (e.g., earlier in the school year vs. towards the end of the school year) could easily be available and might capture different students' motivation.

Implications for Instructional Practice

The findings in this study indicated that not all text features affect readers in the same fashion. Specifically, readers with higher general vocabulary knowledge benefitted more from passages with shorter sentences and more familiar words than their peers with lower vocabulary knowledge, especially when they are asked to recall specific details from a localized section of the text without referring back to the source passage. However, the same was not true with *temporality*: the passages with more time makers helped students with lower vocabulary knowledge while students with greater vocabulary knowledge benefitted more from the low temporality passages.

Based on these findings, it is tempting to speculate that a robust program of vocabulary learning might lay the groundwork for students to "exploit" other affordances, such as the clues that syntax provides about the logical relationships among words (e.g., A is an attribute for B, A and B belong to the same class). However drawing such implication from a single correlational study is risky. Clearly, correlation is not causation; it is just as likely, at least hypothetically, that directly teaching students how to exploit the logical relations among words that syntax indexes may boost vocabulary learning, particularly the incidental learning students engage in while reading or listening to discourse about how the world works. Prudence, therefore, suggests that these explanatory findings among important variables should be used to generate interesting, and highly

plausible, hypotheses for future pedagogical experiments that should be tested experimentally in the ecologically valid manner.

Implications for Measurement

Examining the sources of difficulty has important implications for validity of assessments, item generation, and score interpretation. It plays a key role in determining construct representation by specifying cognitive components that underlie students' performance (Embretson, 1998). In essence, psychometric models that incorporate explanatory cognitive variables on the item side (e.g., LLTM and LLTM+e) decompose item difficulty into cognitive components that are tied to the passage and task features. When successfully modeled with appropriate predictors, these models offer predictive weights that would enable the test developer to control sources of difficulty. Such information is useful for selecting and designing items for future test administration.

The current study revealed that the difficulty of ReadingPlus' Insight Assessment is primarily explained by the passage features such as average words familiarity, syntactic complexity (e.g., sentence length) and temporality that affect text representation rather than the task features related to selecting a response option. This finding appears to be in line with the goal of the ReadingPlus Insight Assessment as a tool to determine students' initial reading level so that they could be placed appropriately in the online instructional program. As such, it is reasonable to assume that the assessment's primary focus is more on students' ability to construct adequate text representation(s) to recall and summarize information from the text, rather than on their verbal and problem solving ability needed for response selection (recall, the latter type of ability appeared to be emphasized more in GRE). Consequently, the findings suggest that scores from the ReadingPlus Insight Assessment reflect students' ability to construct text representation more than their verbal problem-solving skills.

The current study also offers a few suggestions for future item and passage generation for the ReadingPlus Insight Assessment. First, the developer could use sentence length, other syntactic features (e.g., the number of modifiers per noun phrase), and the consistency and amount of time makers, for adjusting the processing difficulty, in addition to the mean log word frequency that is currently used to place the assessment passages into one of the twelve levels. Second, the developer might reconsider whether asking students to recall specific information localized within a sentence is important for the purpose of the assessment, especially when students are not allowed to refer back to the text. Recall that the literal recall questions in the ReadingPlus Insight Assessment were not the easiest among the four item types examined. This item type appears to be most vulnerable to the criticism directed towards the tests that do not allow passage access; they transform an RC test into a test of memory as well as of RC.

Conclusion

This study offered ways to explain variations in item difficulties among reading comprehension questions by simultaneously modeling explanatory variables about the reader, the text, and the task as envisioned in the RAND heuristic of reading comprehension. Importantly, the explanatory item response modeling approach uncovered possible sources of text processing difficulty in the ReadingPlus Insight assessment that differentially affected the comprehension of readers depending on students' general vocabulary knowledge as well as the specific demands of different item types. The findings indicate that a number of factors contribute to the manifestation of RC difficulties in complex ways. Ultimately understanding these complex interactions among the reader, the passage,

and the task will help identifying students with comprehension difficulties and designing targeted interventions that best facilitate their RC development.

Appendix A

Coding Scheme 1: Passage / Question Relations
This coding system is based on Ozuru et al. (2008) and addresses the type of passage comprehension processes that test takers needed to engage, in order to answer the question correctly. There were four levels in this scheme.

Level 1: Text-based question [TE]

- the answer to the question is explicitly stated **within a single sentence in almost verbatim fashion**,
- minimal text processing is required
- The question targets at the comprehension of information explicitly stated within a sentence

Level 2: Restructuring/rewording within a sentence or a paragraph. [RS]

- The target information to answer this type of question is located in the same paragraph, but may cut across a few neighboring sentences within the paragraph, AND
- the target information for the question is **reworded or restructured** (i.e., not verbatim)
- This is a deeper level comprehension than text-base question as it requires of restructuring and/or rewording of a sentence or neighboring sentences in a paragraph.

Level 3: Integration or Bridging question [IB]

- Answering this type of question requires **some degree of integration of information located across multiple paragraphs** from the source passage.
- In other words, an IB question requires across-paragraph integration of information

Level 4: Knowledge-based inference questions [KI]

- In this type of question, the information required to answer a question was **not explicitly stated in the source passage**. The test takers had to make inferences about the situation described in the passage **on the basis of their prior knowledge**

Note. Ozuru et al. (2008) drew on Kintsch's construction integration model of comprehension (Kintsch, 1988, 1998).

Coding Scheme 2: Abstractness of Information Requested by the Question
This coding scheme is based on Mosenthal's (1996) and addresses the abstractness of the information requested by an item. It classifies questions into the following four levels:

Level 1: **Highly concrete** information is asked
- the identification of persons, animals, things, or concrete actions

Level 2: **Somewhat concrete** information is asked
- the identification of amounts, times, or attributes

Level 3: **Somewhat abstractness** information is asked
- Example: identification of manner, goal, purpose, alternative, attempt, or condition, cause, effect, reason, or result;
- Also includes the identification of invisible or intangible actions such as "thinking" or "feeling

Level 4: **Highly abstract** information is asked
- the identification of equivalence, difference, or theme
- the term equivalence in this case refers to highly unfamiliar or low-frequency vocabulary items for which respondents must provide a definition

Coding Scheme 3: the Quality of Distractor Options (falsifiability)

This coding scheme is based on Embretson and Wetzel (1987) and in Ozuru et al (2008). It identifies the number of distractors that could have been explicitly falsified by the content of the passage.

- A distractor was falsifiable if the passage provided explicit textual evidence that the distractor was incorrect.
- A distractor was not falsifiable if the passage had no explicit mention of the distractor
- For each distractor/foil, record whether it is falsifiable or not, and tally up the number of falsifiable distractors at the item level.

Coding Process:

Step 1. Is the answer option **true** based on the source passage?

- If yes – not falsifiable (give a 0 = no)
- If no or not clear – then move to Step 2:

Step 2: Does the passage provide explicit evidence to determine that the option is **not true**?

- If yes – falsifiable (give a 1 = yes)
- If no – not falsifiable (give a 0 = no)

Step 3: Add up the number of falsifiable distractors per question (max possible: 4)

Appendix B

Tables and figures below were generated with the cross validation sample (n=5,273).

Table B-1. Comparison of Text Representation Models

Model	Pseudo R^2	Log-Likelihood	No. of Item Par*	AIC	BIC
M0: Null (Latent regression, LR)	0.000	-38529.08	0	77064.17	77091.05
MS: Saturated (LR-Rasch)	1.000	-37138.07	240	74760.13	76928.35
Text Representation (TR) Models					
M1: Gorin & Embretson	0.443	-37913.44	3	75838.89	75892.64
M2: Lexile	0.504	-37828.62	2	75667.24	75712.03
M3: Coh-Metrix	0.518	-37808.35	8	75638.70	75737.25
M4: TextEvaluator	0.521	-37804.90	8	75631.79	75730.35
M5: Lexile + Coh-Metrix	**0.542**	**-37775.01**	**10**	**75576.01**	**75692.49**
M6: Lexile + TextEvaluator	0.527	-37743.71	10	75616.80	75733.27

Note. Bolded is the best fitting model among the five TR models in the table.
* Number of item parameters estimated. Note there is an additional person parameter (student's vocabulary level) in each model.

Table B-2. Parameter Estimates for Text Representation (TR) Models

	M2 Lexile		M3 Coh-Metrix		M5 Lexile + Coh-Metrix	
	Est.	SE	Est.	SE	Est.	SE
Fixed Effects						
Item						
Mean sent length	.32***	.02			.24***	.04
Log mean word freq	-.13***	.02			-.13***	.03
Narrativity			-.36***	.03	-.10*	.04
Syntactic simplicity			-.44***	.03	-.16***	.05
Word concreteness			-.07***	.02	-.01	.02
Referential cohesion			>.01	.02	-.01	.02
Deep cohesion			.06***	.01	.02	.01
Verb cohesion			-.11***	.01	>-.01	.02
Logical cohesion			-.01	.01	-.04**	.01
Temporal cohesion			-.10***	.01	-.07***	.02
Person						
Vocabulary level	.38***	.02	.33***	.02	.34***	.02
Intercept	-.10***	.01	-.18***	.05	-.22***	.05
Random Effects						
reader variance $\sigma^2_{\epsilon_\theta}$	.44**	.02	.42***	.02	.43***	.02

Note. Est. columns show fixed or random effects in logit, analogous to the standardized regression coefficient (beta weights). SE stands for standard error. Bolded are the significant effects replicated in the cross-validation analysis. * p<.05 ** p<.01 *** p<.001

Table B-3. Comparison of Response Decision Models

Model	Pseudo R^2	Log-Likelihood	No. of Item Par	AIC	BIC
Response Decision Models					
M7: Readability of question/choices	.17	-38293.93	3	76599.86	76653.62
M8: Item type/comprehension process	.09	-38409.14	3	76830.28	76884.04
M9: Abstractness of info asked	.05	-38462.32	3	76936.64	76990.40
M10: Falsifiability of distractors	.02	-38496.89	1	77001.79	77037.63
M11: All RD predictors (M7-M10)	**.27**	**-38147.98**	**10**	**76321.97**	**76438.44**
TR + RD Combined Model					
M12: TR + RD combined (M5+M11 without Falsifiability)	**.58**	**-37719.81**	**20**	**75483.63**	**75680.74**

Table B-4. Parameter Estimates for Response Decision (RD) Models

	M7 Vocab Demand of Item & Ans. Choices		M8 Item Type / Comprehension Process		M11 All RD Predictors	
	Est.	SE	Est.	SE	Est.	SE
Fixed Effects						
Item						
Vocab Demand, Question	.04**	.01			>.01	.01
Vocab Demand, Correct Answer	.17***	.01			.20***	.01
Vocab Demand, Distractors	.06***	.01			.04**	.01
Item type (ref = text-base)						
reword/restructure			-.03	.03	-.14***	.05
bridging			-.09*	.04	-.15***	.06
knowledge-base			.24***	.03	.14***	.05
Abstractness of Info (ref=highly concrete)						
somewhat concrete					.29***	.04
somewhat abstract					.13***	.04
highly abstract					.13**	.04
Falsifiability					>-.01	.02
Person						
Vocabulary level	.30***	.02	.26***	.02	.29***	.02
Intercept	-.21***	.01	-.14	.03	-.10	.04
Random Effects						
reader variance $\sigma^2_{\epsilon_\theta}$	.44***	.02	.44***	.02	.46***	.02

* p<0.05 ** p<0.01 *** p<0.001

Table B-5. Parameter Estimates for TR and RD Combined Model (M12)

	M12 TR + RD combined	
	Est.	SE
Fixed Effects		
Text Representation (TR)		
Mean sent length	0.25***	0.04
Log mean word freq	-0.10***	0.03
Narrativity	-0.11*	0.05
Syntactic simplicity	-0.15**	0.05
Word concreteness	0.04	0.02
Referential cohesion	>0.01	0.02
Deep cohesion	-0.01	0.01
Verb cohesion	>-0.01	0.02
Logical cohesion	0.04**	0.01
Temporality	-0.06***	0.02
Response Decision (RD)		
Vocab Demand, Question	-0.01	0.012
Vocab Demand, Correct Ans.	0.05***	0.012
Vocab Demand, Distractors	-0.03**	0.011
Item Type (ref = text-base)		
restructure	-0.07	0.039
integrate	-0.08	0.045
knowledge-base	0.06	0.041
Abstractness of Info (ref=highly concrete)		
somewhat concrete	0.23***	0.035
somewhat abstract	0.14***	0.031
highly abstract	0.11**	0.037
Reader		
Vocabulary level	0.34***	0.02
Intercept	-0.08**	0.07
Random Effects		
Reader variance $\sigma^2_{\epsilon_\theta}$	0.43***	0.02

* p<0.05 ** p<0.01 *** p<0.001

Table B-6. Comparison of Interaction Models by Pseudo R2, AIC, and BIC

Model	Pseudo R^2	Δ^+ Pseudo R^2	Log-Likelihood	No. of Item Par+	AIC	BIC
TR + RD Combined Model						
M12: TR + RD combined	.582	--	-37719.81	20	75483.63	**75680.74**
Text-Reader Interaction Models						
M12: M12 + MSL×voc	.586	.004***	-37713.81	21	75473.62	75679.69
M13: M12 + MLWF×voc	.589	.007***	-37709.39	21	75464.78	75670.85
M14: M12 + Synt ×voc	.584	.002*	-37716.66	21	75479.32	75685.39
M15: M12 + Temp×voc	.582	.000	-37719.58	21	75485.17	75691.24
Text-Task Interaction Models						
M16: M12 + MSL×IType	.588	.006***	-37710.57	23	75471.13	75695.12
M17: M12 + MLWF×IType	.596	.014***	-37700.00	23	75450.00	75673.99
M18: M12 + Synt×IType	.585	.003*	-37715.83	23	75481.66	75705.65
M19: M12 + Temp×IType	.606	.024***	-37685.71	23	75421.42	75645.41
Text-Reader-Task Interaction Models						
M20 M12 + MSL×IType×voc	**.619**	**.037*** **	-37668.72	30	**75401.44**	**75688.14**
M21: M12 + MLWF×IType×voc	.609	.027***	-37681.52	30	75427.04	75713.75
M22: M12 + Synt× IType ×voc	.604	.022***	-37689.07	30	75442.14	75728.85
M23: M12 + Temp×IType×voc	.617	.035***	**-37670.22**	30	75404.45	75691.15

+ Change in Pseudo R^2 from the main-effects only model (M11) * p<0.05 ** p<0.01 *** p<0.001

Note. Bolded are the values that are the most extreme for the model comparison purpose: the largest for pseudo R^2 and change in pseudo R^2 and the smallest for log-likelihood, AIC, and BIC.

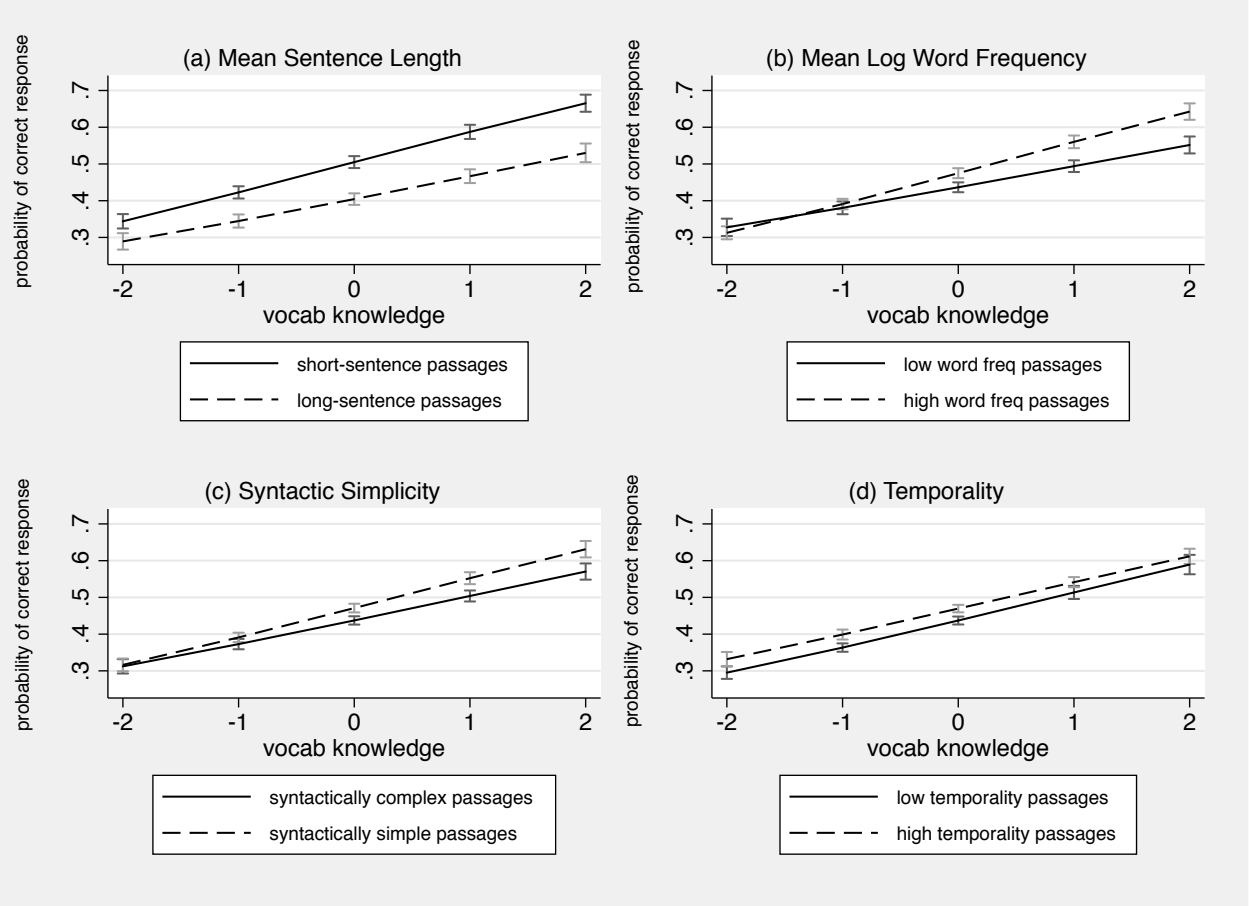

Figure B-1. Four panels of line plots, each depicting interactions between general vocabulary knowledge and one of the four text features: (a) mean sentence length, (b) mean log word frequency, (c) syntactic simplicity, and (d) temporality, after controlling for all other text and item variables. All but the panel (d) shows the modification of the text effects by general vocabulary knowledge, as evident with the widening of the gap between the two lines. Two levels of text feature variables were set at 1 SD above and 1 SD below their respective means.

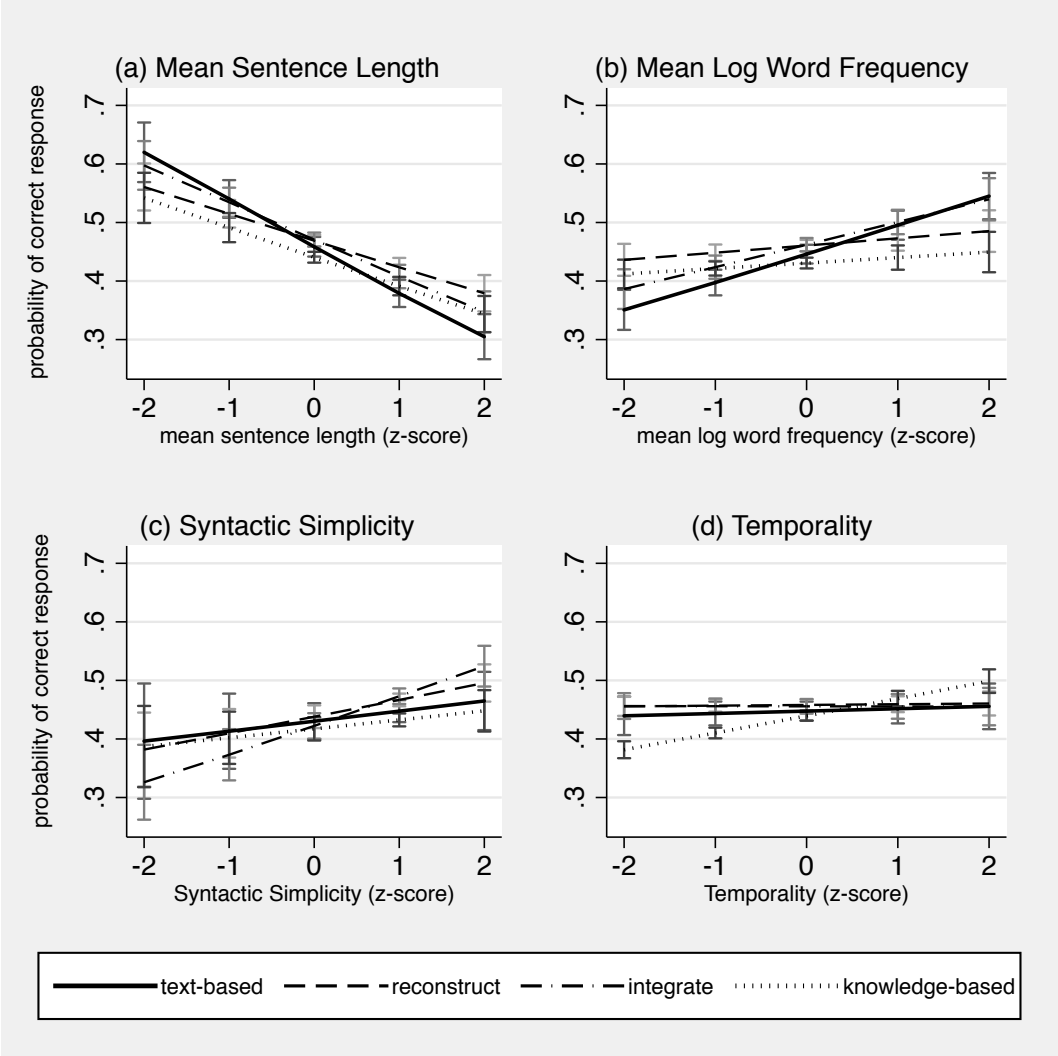

Figure B-2. Four panels of line plots, each depicting interactions between the item type and one of the four text features: (a) mean sentence length, (b) mean log word frequency, (c) syntactic simplicity, and (d) temporality, after controlling for all other text and task variables in the model. All but the panel (c) shows the modification of the text feature by the item type.

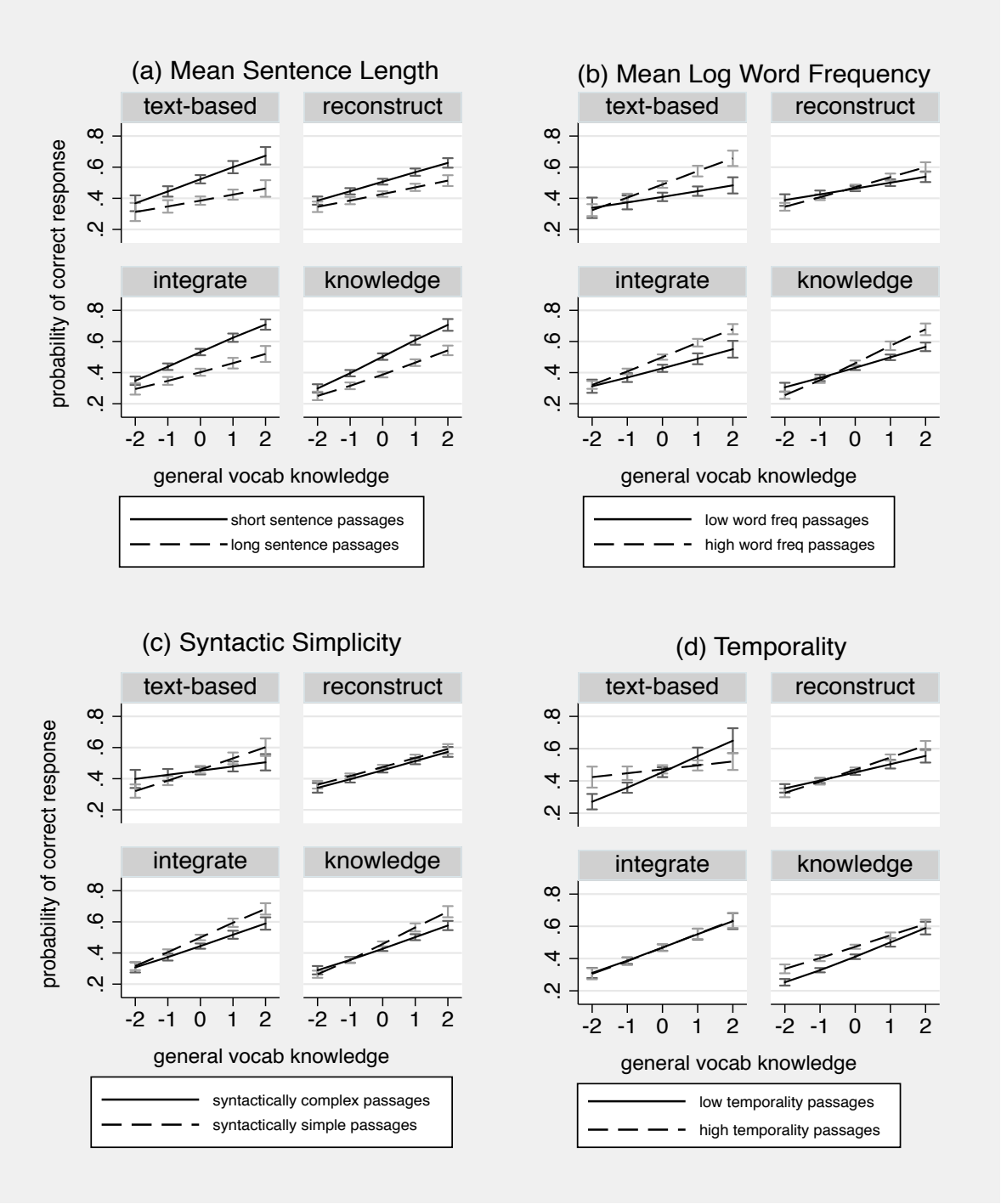

Figure B-3.Four panels of line plots, each depicting three-way interactions among reader's general vocabulary knowledge, item type, and one of the four text features: (a) the mean sentence length, (b) the mean log word frequency, (c) syntactic simplicity, and (d) temporality, after controlling for all other text and item variables. All but the panel (d) shows that the simultaneous effects of temporality and the item type were moderated by the reader's general vocabulary knowledge.

References

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal*, *76*(4), 468–479. https://doi.org/10.1111/j.1540-4781.1992.tb05394.x

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC-19*(6): 716-723

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Amendum, S. J., Conradi, K., & Hiebert, E. (2018). Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students' reading fluency and comprehension. *Educational Psychology Review*, *30*(1), 121–151. https://doi.org/10.1007/s10648-017-9398-2

Anderson, R. C., & Davison, A. (1986). Conceptual and empirical basaes of readability formulas. In A. DAVISON & G. M. GREEN (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered* (pp. 23–54). Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, Richard C. (1972). How to Construct Achievement Tests to Assess Comprehension. *Review of Educational Research*, *42*(2), 145–170. https://doi.org/10.3102/00346543042002145

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education*, *16*(3), 363–383. https://doi.org/10.1007/BF03173188

Bachman, L. F. (2006). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 535–556. https://doi.org/10.2307/3586277

Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly*, *1*, 79–132.

Bormuth, J. R. (1969). *Development of readability analysis (Final Report, Project No. 7-0052, Contract No. OEC-3-7-070052-0326)* Washington, D.C.: Office of Education, Bureau of Research, U.S. Department of Health, Education, and Welfare.

Bormuth, John R. (1968). Cloze test readability: Criterion reference scores. *Journal of Educational Measurement*, *5*(3), 189–196. https://doi.org/10.1111/j.1745-3984.1968.tb00625.x

Bouwmeester, S., van Rijen, E. H. M., & Sijtsma, K. (2011). Understanding phoneme segmentation performance by analyzing abilities and word properties. *European Journal of Psychological Assessment*, *27*(2), 95–102. https://doi.org/10.1027/1015-5759/a000049

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25. https://doi.org/10.1080/01621459.1993.10594284

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89–118. https://doi.org/10.1080/08957340801926086

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, *50*(2), 204–226. https://doi.org/10.1111/jedm.12011

Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, *83*(3), 329–345. https://doi.org/10.1037/0022-0663.83.3.329

Broek, P. Van Den, & Espin, C. a. (2012). Connecting theory and assessment: Measuring individual differences in Reading Comprehens. *School Psychology Review*, *41*(3), 315–325.

Bruce, B., & Rubin, A. (1988). Readability formulas: Matching tool and task. In *Linguistic complexity and text comprehension: Readability issues reconsidered* (pp. 5–22).

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent

prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*(1), 31–42. https://doi.org/10.1037/0022-0663.96.1.31
Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.
Chall, J. S., Bissex, G. L., Conard, S. S., & Harris-Sharples, S. H. (1996). *Qualitative assessment of text difficulty: A practical guide for teachers and writers.* Cambridge, MA: Brookline.
Coleman, D., & Pimentel, S. (2012). *Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12*. Retrieved June 28, 2019 from http://www.corestandards.org/assets/Publishers_Criteria_for_3-12.pdf
Cunningham, J. W., & Mesmer, H. A. (2014). Quantitative measurement of text difficulty: What's the use? *The Elementary School Journal*, *115*(2), 255–269. https://doi.org/https://doi.org/10.1086/678292
Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*(3), 277–299. https://doi.org/10.1207/s1532799xssr1003_5
De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559. https://doi.org/10.1007/s11336-008-9092-x
De Boeck, P., & Wilson, M. (2004). *Explanatory item response theory models: A generalized linear and nonlinear approach*. New York, NY: Springer.
Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*(1), 1–19. https://doi.org/10.1007/s11336-013-9342-4
Drum, P. a, Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, *16*(4), 486–514.
Duke, N. K. (2005). Comprehension of what for what: Comprehension as a non-unitary consruct. In *Current issues in reading comprehension and assessment* (pp. 93–104). Mahwah, NJ: Lawrence Erlbaum Associates.
Embretson, S. E. (1983). An incremental fit index for the linear logistic latent trait model. In *annual meeting of the Psychometric Society*. Los Angeles, CA.
Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396. https://doi.org/10.1037/1082-989X.3.3.380
Embretson, Susan E, & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*(2), 175–193. https://doi.org/10.1177/014662168701100207
Englert, C. S., Raphael, T. E., Anderson Helene M. Anthony, L. M., & Stevens, D. D. (1991). Making Strategies and Self-Talk Visible: Writing Instruction in Regular and Special Education Classrooms. *American Educational Research Journal*, *28*(2), 337–372. https://doi.org/10.3102/00028312028002337
Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6
Fry, E. (1977). Fry's readability graph: Clarification, validity, and extension to level 17. *Journal of Reading*, *21*, 242–252.
Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment*, *31*(1), 16–28. https://doi.org/10.1177/0734282912451971

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.
Gilbert, J. K., Compton, D. L., & Kearns, D. M. (2011). Word and person effects on decoding accuracy: A new look at an old question. *Journal of Educational Psychology*, *103*(2), 489–507. https://doi.org/10.1037/a0023001
Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*(4), 351–373. https://doi.org/10.1111/j.1745-3984.2005.00020.x
Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394–411. https://doi.org/10.1177/0146621606288554
Graesser, A.C, McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234. https://doi.org/10.3102/0013189X11413260
Graesser, Arthur C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*(2), 371–398. https://doi.org/10.1111/j.1756-8765.2010.01081.x
Guthrie, J. T., Wigfield, A., Barbosa, P., Perencevich, K. C., Taboada, A., Davis, M. H., … Tonks, S. (2004). Increasing reading comprehension and engagement through concept-oriented reading instruction. *Journal of Educational Psychology*, *96*(3), 403–423. https://doi.org/10.1037/0022-0663.96.3.403
Harris, D. J. (2007). Practical issues in vertical scaling. In *Linking and Aligning Scores and Scales* (pp. 233–251). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-49771-6_13
Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, *72*(4), 665–686. https://doi.org/10.1177/0013164411430707
Higgs, K., Magliano, J. P., Vidal-Abarca, E., Martínez, T., & McNamara, D. S. (2017). Bridging skill and task-oriented reading. *Discourse Processes*, *54*(1), 19–39. https://doi.org/10.1080/0163853X.2015.1100572
Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, *2*, 127–160.
Hua, A. N., & Keenan, J. M. (2014). The role of text memory in inferencing and in comprehension deficits. *Scientific Studies of Reading*, *18*(6), 415–431. https://doi.org/10.1080/10888438.2014.926906
Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In *Explanatory Item Response Models* (pp. 189–212). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-3990-9_6
Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn & Bacon.
Kamil, M. L. (2001). Comments on Lexile framework. In S. White & J. Clement (Eds.), *Assessing the Lexile Framework. Results on a Panel Meeting* (pp. 22–26). National Center for Education Statistics.
Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess : Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 37–41.

Kendeou, P., van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology*, *101*(4), 765–778. https://doi.org/10.1037/a0015956

Kim, Y.-S., Petscher, Y., Foorman, B. R., & Zhou, C. (2010). The contributions of phonological awareness and letter-name knowledge to letter-sound acquisition—a cross-classified multilevel model approach. *Journal of Educational Psychology*, *102*(2), 313–326. https://doi.org/10.1037/a0018449

Kincaid, J. P., Fishburne, L. R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. In *Research Branch Report* (pp. 8–75). Memphis: Chief of Naval Technical Training: Naval Air Station.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kintsch, W. (1988). The role of knowledge in discourse processing: A construction-incrementation model. *Psychological Review*, *95*(2), 163–182. https://doi.org/http://dx.doi.org/10.1037/0033-295X.95.2.163

Kintsch, Walter, & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394.

Kirby, J. R., Ball, A., Geier, B. K., Parrila, R., & Wade-Woolley, L. (2011). The development of reading interest and its relation to reading ability. *Journal of Research in Reading*, *34*(3), 263–280. https://doi.org/10.1111/j.1467-9817.2010.01439.x

Klare, G. R. (1984). Readability. In P. D. Pearson, R. Barr, & M. L. Kamil (Eds.), *Handbook of Reading Research* (pp. 681–744). New York, NY: Longman.

Klare, George R. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press.

Kobayashi, M. (2003). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal*, *86*(4), 571–586. https://doi.org/10.1111/1540-4781.00162

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.

Koslin, B. L., Zeno, S., & Koslin, S. (1987). *The DRP: An effective measure in reading*. New York, NY.

Kulesz, P. A., Francis, D. J., Barnes, M. A., & Fletcher, J. M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology*, *108*(8), 1078–1097. https://doi.org/10.1037/edu0000126

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, *22*(3), 249–264. https://doi.org/10.3102/10769986022003249

Magliano, J. P., Todaro, S., Millis, K., Wiemer-Hastings, K., Kim, H. J., & McNamara, D. S. (2005). Changes in reading strategies as a function of reading training: A comparison of live and computerized training. *Journal of Educational Computing Research*, *32*(2), 185–208. https://doi.org/10.2190/1LN8-7BQE-8TN0-M91L

Markus, K., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2001). *Generalized, linear, and mixed models*.

New York, NY: Wiley.
McNamara, D.S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge, MA: Cambridge University Press.
McNamara, Danielle S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *55*(1), 51–62. https://doi.org/10.1037/h0087352
McNamara, Danielle S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1–43. https://doi.org/10.1207/s1532690xci1401_1
McNamara, Danielle S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247–288. https://doi.org/10.1080/01638539609544975
McNamara, Danielle S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes*, *47*(4), 292–330. https://doi.org/10.1080/01638530902959943
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington DC: American Council on Education and National Council on Measurement in Education.
Miller, A. C., Davis, N., Gilbert, J. K., Cho, S. J., Toste, J. R., Street, J., & Cutting, L. E. (2014). Novel approaches to examine passage, student, and question effects on reading comprehension. *Learning Disabilities Research and Practice*, *29*(1), 25–35. https://doi.org/10.1111/ldrp.12027
Miller, G. A., & Gildea, P. M. (1987). How children learn words. *Scientific American*, *257*(3), 94–99. https://doi.org/10.1038/scientificamerican0987-94
Mislevy, R. (1987). Exploiting auxiliary infornlation about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11*(1), 81–91.
Mosenthal, P. B. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology*, *88*(2), 314–332. https://doi.org/10.1037/0022-0663.88.2.314
National Assessment Governing Board. (2015). *Reading Framework for the 2015 National Assessment of Educational Progress*. Washington, DC.
National Governors Association[NGA], & Council of Chief State School Officers[CCSSO]. (2010). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved June 2, 2019 from http://www.corestandards.org/
Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value of grade levels and student performance. Report to the Gates foundation*. New York. Retrieved May 13, 2019 from http://www.ccsso.org/Documents/2012/Measures ofText
O'Reilly, T., & Mcnamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, *43*(2), 121–152. https://doi.org/10.1080/01638530709336895
Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence From a four-year longitudinal study. *Scientific Studies of Reading*, *16*(2), 91–121. https://doi.org/10.1080/10888438.2010.529219
Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: the passage or the question? *Behavior Research Methods*, *40*(4),

1001–1015. https://doi.org/10.3758/BRM.40.4.1001
Patz, R. J., & Hanson, B. A. (2002). Psychometric issues in vertical scaling. In *Paper presented at the annual meeting of the National Council on Measurement in Education*. New Orleans, LA.
Patz, Richard J., & Yao, L. (2006). Vertical scaling: Statistical models for measuring growth and achievement. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, 26(6),* 955–975. Amsterdam: North Holland. https://doi.org/10.1016/S0169-7161(06)26030-9
Pearson, P. D., & Cervetti, G. N. (2015). Fifty years of reading comprehension theory and practice. In P. D. Pearson & E. H. Hiebert (Eds.), *Research-Based Practices for Teaching Common Core Literacy* (pp. 1–40). New York: Teachers College, Columbia University.
Pearson, P. D., Hansen, J., & Gordon, C. (1979). The effect of background knowledge on young children's comprehension of explicit and implicit information. *Journal of Reading Behavior*, *11*(3), 201–209. https://doi.org/10.1080/10862967909547324
Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730
Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687
RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA.
Rapp, D. N., Broek, P. van den, McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for Research and Intervention. *Scientific Studies of Reading*, *11*(4), 289–312. https://doi.org/10.1080/10888430701530417
Rinker, T. (2013). qdap: Quantitative discourse analysis package. Retrieved March 1, 2019 from http://trinker.github.io/qdap/
Rupp, A. a, Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing*, *23*(4), 441–474. https://doi.org/10.1191/0265532206lt337oa
Santi, K. L., Kulesz, P. a., Khalaf, S., & Francis, D. J. (2015). Developmental changes in reading do not alter the development of visual processing skills: an application of explanatory item response models in grades K-2. *Frontiers in Psychology*, *6*(116), 1–16. https://doi.org/10.3389/fpsyg.2015.00116
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. Retrieved from http://projecteuclid.org/euclid.aos/1176344136
Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential xomprehension. *Reading Research Quarterly*, *17*(2), 229. https://doi.org/10.2307/747485
Sheehan, K. M. (2016). *A review of evidence presented in support of three key claims in the validity argument for the TextEvaluator® text analysis tool (Research Report No. RR-16-12)*. Parsippany, NJ: John Wiley & Sons, Ltd. https://doi.org/10.1002/ets2.12100
Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool : Helping teachers and test developers select texts for use in instruction and assessment. *Elementary School Journal*, *115*(2), 184–209. https://doi.org/https://doi.org/10.1086/678294
Smith, D. R., Stenner, A. J., Horabin, I., & Smith, M. (1989). *The Lexile scale in theory and practice. Final report*. Durham, NC (ED307577).
Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, *56*(1), 72–110. https://doi.org/10.3102/00346543056001072
Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are lexile text

measures? *Journal of Applied Measurement*, *7*(3), 307–322.
Stenner, A. Jackson, Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*(4), 305–316. https://doi.org/10.1111/j.1745-3984.1983.tb00209.x
Stenner, A Jackson. (1996). Measuring reading comprehension with the Lexile framework. In *Paper presented at the California Comparability Symposium*. Washington, DC.
Stenner, J., Fisher, W. P., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology*, *4*(August), 536. https://doi.org/10.3389/fpsyg.2013.00536
Svetina, D., Gorin, J. S., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing*, *11*(1), 1–23. https://doi.org/10.1080/15305058.2010.518261
Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading*, *10*(4), 381–398. https://doi.org/10.1207/s1532799xssr1004_3
Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401
Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice*, *15*(1), 55–64. https://doi.org/10.1207/SLDRP1501_6
Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*(4), 333–344. https://doi.org/10.1177/014662168601000402
Van den Noortgate, W., & Paek, I. (2004). Person regression models. In *Explanatory Item Response Models* (pp. 167–187). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-3990-9_5
Wang, W., Cheng, Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement*, *65*(1), 5–27. https://doi.org/10.1177/0013164404268676
Wu, M., Adams, R., & Wilson, M. (1998). *ACER ConQuest: Generalised item response modelling software*. Retrieved February 14, 2019 from http://works.bepress.com/ray_adams/25/
Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162
Zwinderman, A. H. (1991). A generalized rasch model for manifest predictors. *Psychometrika*, *56*(4), 589–600. https://doi.org/10.1007/BF02294492